Metric entropy in competitive on-line prediction

Vladimir Vovk
vovk@cs.rhul.ac.uk
http://vovk.net

February 1, 2008

Abstract

Competitive on-line prediction (also known as universal prediction of individual sequences) is a strand of learning theory avoiding making any stochastic assumptions about the way the observations are generated. The predictor's goal is to compete with a benchmark class of prediction rules, which is often a proper Banach function space. Metric entropy provides a unifying framework for competitive on-line prediction: the numerous known upper bounds on the metric entropy of various compact sets in function spaces readily imply bounds on the performance of on-line prediction strategies. This paper discusses strengths and limitations of the direct approach to competitive on-line prediction via metric entropy, including comparisons to other approaches.

1 Introduction

A typical result of competitive on-line prediction says that, for a given benchmark class of prediction strategies, there is a prediction strategy that performs almost as well as the best prediction strategies in the benchmark class. For simplicity, in this paper the performance of a prediction strategy will be measured by the cumulative squared distance between its predictions and the true observations, assumed to be real (occasionally complex) numbers. Different methods of competitive on-line predictions (such as Gradient Descent, following the perturbed leader, strong and weak aggregating algorithms, defensive forecasting, etc.) tend to have their narrow "area of expertise": each works well for benchmark classes of a specific "size" but is not readily applicable to classes of a different size.

In this paper we will apply a simple general method based on metric entropy to benchmark classes of a wide range of sizes. Typically, this method does not give optimal results, but its results are often not much worse than those given by specialized methods, especially for benchmark classes that are not too massive. Since the method is almost universally applicable, it sheds new light on the known results.

Another disadvantage of the metric entropy method is that it is not clear how to implement it efficiently, whereas many other methods are computationally very efficient. Therefore, the results obtained by this method are only a first step, and we should be looking for other prediction strategies, both computationally more efficient and having better performance guarantees.

We start, in §2, by stating a simple asymptotic result about the existence of a universal prediction strategy for the class of continuous prediction rules. The performance of the universal strategy is in the long run as good as the performance of any continuous prediction rule, but we do not attempt to estimate the rate at which the former approaches the latter. This is the topic of the following section, §3, where we establish general results about performance guarantees based on metric entropy. For example, in the simplest case where the benchmark class $\mathcal{F}$ is a compact set, the performance guarantees become weaker as the metric entropy of $\mathcal{F}$ becomes larger.

The core of the paper is organized according to the types of metric compacts pointed out by Kolmogorov and Tikhomirov in [27] (§3). Type I compacts have metric entropy of order $\log \frac{1}{\epsilon}$; this case corresponds to the finite-dimensional benchmark classes and is treated in §4. Type II, with the typical order $\log^M \frac{1}{\epsilon}$, contains various classes of analytic functions and is dealt with in §5. The key §6 deals with perhaps the most important case of order $\left(\frac{1}{\epsilon}\right)^\gamma$; this includes, e.g., Besov classes. The classes of type IV, considered in §7, have metric entropy that grows even faster.

In §§4–7 the benchmark class is always given. In §9 we ask the question of how prediction strategies competitive against various benchmark classes compare to each other. The previous section, §8, prepares the ground for this. The concluding section, §10, lists several directions of further research.

There is no real novelty in this paper; I just apply known results about metric entropy to competitive on-line prediction. I hope it will be useful as a survey.

2 Simple asymptotic result

Throughout the paper we will be interested in the following prediction protocol (or its modifications):

ON-LINE REGRESSION PROTOCOL
FOR $n = 1, 2, \ldots$:
  Reality announces $x_n \in \mathbf{X}$.
  Predictor announces $\mu_n \in \mathbb{R}$.
  Reality announces $y_n \in [-Y, Y]$.
END FOR.

At the beginning of each round n Predictor is given some *signal* x_n that might be helpful in predicting the following *observation* y_n, after which he announces his *prediction* μ_n. The signal is taken from the *signal space* $\mathbf{X}$, the observations

are real numbers known to belong to a fixed interval $[-Y, Y]$, $Y > 0$, and the predictions are any real numbers (later this will also be extended to complex numbers). The error of prediction is always measured by the quadratic loss function, so the loss suffered by Predictor on round n is $(y_n - \mu_n)^2$. It is clear that it never makes sense for Predictor to choose predictions outside $[-Y, Y]$, but the freedom to go outside $[-Y, Y]$ might be useful when the benchmark class is not closed under truncation.

Remark Competitive on-line prediction uses a wide range of loss functions $\lambda(y_n, \mu_n)$. The quadratic loss function $\lambda(y_n, \mu_n) := (y_n - \mu_n)^2$ belongs to the class of "mixable" loss functions, which are strictly convex in the prediction μ_n in a fairly strong sense. Such loss functions allow the strongest performance guarantees (using, e.g., the "aggregating algorithm" of [42], which we will call the strong aggregating algorithm). If the loss function is convex but not strictly convex in the prediction, the performance guarantees somewhat weaken (and can be obtained using, e.g., the weak aggregating algorithm of [23]; for a review of earlier methods, see [12]). When the loss function is not convex in the prediction, it can be "convexified" by using randomization ([12], Chapter 4).

A *prediction rule* is a function $F : \mathbf{X} \to \mathbb{R}$. Intuitively, F plays the role of the strategy for Predictor that recommends prediction $F(x_n)$ after observing signal $x_n \in \mathbf{X}$; such strategies are called Markov prediction strategies in [48]. We will be interested in benchmark classes consisting of only Markov prediction strategies; in practice, this is not as serious a restriction as it might appear: it is usually up to us what we want to include in the signal space $\mathbf{X}$, and we can always extend $\mathbf{X}$ by including, e.g., some of the previous observations and signals.

Our first result states the existence of a strategy for Predictor that asymptotically dominates every continuous prediction rule (for much stronger asymptotic results, see [46, 47, 48, 49]).

Theorem 1 *Let $\mathbf{X}$ be a metric compact. There exists a strategy for Predictor that guarantees*

$$\limsup_{N\to\infty} \left(\frac{1}{N} \sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N} \sum_{n=1}^{N} (y_n - F(x_n))^2 \right) \le 0 \tag{1}$$

for each continuous prediction rule F.

Any strategy for Predictor that guarantees (1) for each continuous $F : \mathbf{X} \to \mathbb{R}$ will be said to be *universal* (or, more fully, *universal for $C(\mathbf{X})$*). Theorem 1, asserting the existence of universal prediction strategies, will be proved at the end of this section.

Aggregating algorithm

This subsection will introduce the main technical tool used in this paper, an aggregating algorithm (in fact intermediate between the strong aggregating algorithm of [42] and the weak aggregating algorithm of [23]). For future use in

§5, we will allow the observations to belong to the Euclidean space $\mathbb{R}^m$ (in fact, we will only be interested in the cases $m = 1$ and $m = 2$). Correspondingly, we allow predictions in $\mathbb{R}^m$ and extend the notion of prediction rule allowing values in $\mathbb{R}^m$.

The l_2 norm in $\mathbb{R}^m$ will be denoted $\|\cdot\|_2$ or simply $\|\cdot\|$; in later sections we will also use l_p norms $\|\cdot\|_p$ for $p \neq 2$. Reality is constrained to producing observations in the ball YU_m in $\mathbb{R}^m$ with radius Y and centred at 0; U_V is our general notation for the closed unit ball $\{v \in V \mid \|v\| \leq 1\}$ centred at 0 in a Banach space V, and we abbreviate $U_m := U_{\mathbb{R}^m}$.

Lemma 1 *Let $F_1, F_2, \ldots$ be a sequence of $\mathbb{R}^m$-valued prediction rules assigned positive weights $w_1, w_2, \ldots$ summing to 1. There is a strategy for Predictor producing $\mu_n \in YU_m$ that are guaranteed to satisfy, for all $N = 1, 2, \ldots$ and all $i = 1, 2, \ldots$,*

$$\sum_{n=1}^{N} \|y_n - \mu_n\|^2 \leq \sum_{n=1}^{N} \|y_n - F_i(x_n)\|^2 + 8Y^2 \ln \frac{1}{w_i}. \tag{2}$$

For $m = 1$ the constant $8Y^2$ in (2) can be improved to $2Y^2$ (cf. [42], Lemma 2 and the line above Remark 3), and it is likely that this is also true in general. In this paper we, however, do not care about multiplicative constants (and usually even do not give them explicitly in the statements of our results; the reader can always extract them from the proofs).

Inequality (2) says that Predictor's total loss does not exceed the total loss suffered by an alternative prediction strategy plus a *regret term* ($8Y^2 \ln \frac{1}{w_i}$ in the case of (2)); we will encounter many such inequalities in the rest of this paper.

Proof of Lemma 1 Let $\eta := \frac{1}{8Y^2}$, $\beta := e^{-\eta}$, and P_0 be the probability measure on $\{1, 2, \ldots\}$ assigning weight w_i to each $i = 1, 2, \ldots$. Lemma 1 and Remark 3 of [42] imply that it suffices to show that the function $\beta^{\|y-\mu\|^2}$ is concave in $\mu \in YU_m$ for each fixed $y \in YU_m$ (this idea goes back to Kivinen and Warmuth [25]). Furthermore, it suffices to show that the function

$$\beta^{\|a+bt\|^2} = e^{-\eta\left(\|a\|^2 + 2\langle a,b\rangle t + \|b\|^2 t^2\right)}$$

is convex in $t \in [0, 1]$ for any a and b such that a and $a + b$ belong to the ball $2YU_m$ of radius $2Y$ centred at 0. Taking the second derivative, we can see that we need to show

$$2\eta\left(\langle a, b\rangle + \|b\|^2 t\right)^2 \leq \|b\|^2.$$

By the convexity of the function $(\cdot)^2$, it suffices to establish the last inequality for $t = 0$,

$$2\eta\left(\langle a, b\rangle\right)^2 \leq \|b\|^2, \tag{3}$$

and $t = 1$,

$$2\eta\left(\langle a, b\rangle + \|b\|^2\right)^2 \leq \|b\|^2. \tag{4}$$

Inequality (3) follows, for $\eta \le \frac{1}{8Y^2}$, from

$$2\eta\big(\langle a,b\rangle\big)^2 \le 2\eta \|a\|^2 \|b\|^2 \le 8\eta Y^2 \|b\|^2 \le \|b\|^2 .$$

In the case of (4), it is clear that we can replace a by its projection onto the direction of b and so assume $a = \lambda b$ for some $\lambda \in \mathbb{R}$. Therefore, (4) becomes

$$2\eta(1+\lambda)^2 \|b\|^4 \le \|b\|^2 .$$

The last inequality, equivalent to $2\eta \|(1+\lambda)b\|^2 \le 1$, immediately follows from the fact that $(1+\lambda)b = a + b$ belongs to $2YU_m$. ∎

The proof of Lemma 1 exhibits an explicit strategy for Predictor guaranteeing (2); we will refer to this strategy as the *AA mixture* of $F_1, F_2, \ldots$ (with weights $w_1, w_2, \ldots$).

Proof of Theorem 1

Theorem 1 follows immediately from the separability of the function space $C(\mathbf{X})$ of continuous real-valued functions on $\mathbf{X}$ ([19], Corollary 4.2.18). Indeed, we can choose a dense sequence $F_1, F_2, \ldots$ of prediction rules in $C(\mathbf{X})$ and take any positive weights w_i summing to 1. Let F be any continuous prediction rule; without loss of generality, $F : \mathbf{X} \to [-Y, Y]$. For any $\epsilon > 0$, the AA mixture clipped to $[-Y, Y]$ will satisfy

$$\begin{aligned}
&\limsup_{N\to\infty} \left(\frac{1}{N}\sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N}\sum_{n=1}^{N} (y_n - F(x_n))^2 \right) \\
&\qquad \le \limsup_{N\to\infty} \left(\frac{1}{N}\sum_{n=1}^{N} (y_n - \mu_n)^2 - \frac{1}{N}\sum_{n=1}^{N} (y_n - F_i(x_n))^2 \right) + 4Y\epsilon \\
&\qquad\qquad \le \limsup_{N\to\infty} \frac{8Y^2 \ln \frac{1}{w_i}}{N} + 4Y\epsilon \le 4Y\epsilon,
\end{aligned}$$

where i is such that F_i is ϵ-close to F in $C(\mathbf{X})$. Since this holds for any $\epsilon > 0$, the proof is complete.

3 Performance guarantees based on metric entropy: general results

In the rest of the paper we will be assuming, without loss of generality, that $Y = 1$. To recover the case of a general $Y > 0$, the universal constants C in Theorems 2–4 (and their corollaries) below should be replaced by CY^2 and all the norms should be divided by Y. The constants C in those theorems are not too large (of order 10 according to the proofs given, but no effort has been made to optimize them).

We will consider three types of non-asymptotic versions of Theorem 1, corresponding to Theorems 2–4 of this section. In the first type the benchmark class $\mathcal{F}$ is a metric compact, and we can guarantee that Predictor's loss over the first N observations does not exceed the loss suffered by the best prediction rule in the benchmark class plus a regret term of $o(N)$, the rate of growth of the regret term depending on the metric entropy of $\mathcal{F}$. In the second type $\mathcal{F}$ is a Banach function space on $\mathbf{X}$ whose unit ball $U_{\mathcal{F}}$ is a compact (in metric $C(\mathbf{X})$) subset of $C(\mathbf{X})$. In this case it is impossible to have the same performance guarantees; Predictor will need a start (given in terms of their norm in the Banach space) on remote prediction rules. Results of this type can be easily obtained from results of the first type. In the third type the benchmark class consists of all continuous prediction rules; such results can be obtained from results of the second type for "universal" Banach spaces, i.e., Banach spaces that are dense subsets of $C(\mathbf{X})$.

Compact benchmark classes

Let A be a compact metric space. The *metric entropy* $\mathcal{H}_\epsilon(A)$, $\epsilon > 0$, is defined to be the binary logarithm $\log N$ of the minimum number of elements $F_1, \dots, F_N \in A$ that form an ϵ-net for A (in the sense that for each $F \in A$ there exists $i = 1, \dots, N$ such that F and F_i are ϵ-close in A). The requirement of compactness of A ensures that $\mathcal{H}_\epsilon(A)$ is finite for each $\epsilon > 0$.

Remark There are four main variations on the notion of metric entropy as defined in [27]; our definition corresponds to Kolmogorov and Tikhomirov's relative ϵ-entropy $\mathcal{H}_\epsilon^A(A)$. In general, relative ϵ-entropy $\mathcal{H}_\epsilon^R(A)$ can be defined for any metric space R containing A as a subspace (in our applications we would take $R := C(\mathbf{X})$). The other two variations are the absolute ϵ-entropy $\mathcal{H}_\epsilon^{\text{abs}}(A)$ (denoted simply $\mathcal{H}_\epsilon(A)$ by Kolmogorov and Tikhomirov; it was introduced by Pontryagin and Shnirel'man [31] in 1932, without taking the binary logarithm and using ϵ in place of Kolmogorov and Tikhomirov's 2ϵ) and the ϵ-capacity $\mathcal{E}_\epsilon(A)$. All four notions were studied by Kolmogorov, his students (Vitushkin, Erokhin, Tikhomirov, Arnol'd), and Babenko in the 1950s, and their results are summarized in [27]. It is always true that, in our notation,

$$\mathcal{E}_{2\epsilon}(A) \le \mathcal{H}_\epsilon^{\text{abs}}(A) \le \mathcal{H}_\epsilon^R(A) \le \mathcal{H}_\epsilon(A) \le \mathcal{E}_\epsilon(A) \tag{5}$$

([27], Theorem IV). All results in [27] can be applied to all elements of the chain (5), and in principle we can use any of the four notions; our choice of $\mathcal{H}_\epsilon(A) = \mathcal{H}_\epsilon^A(A)$ is closest to the notion of entropy numbers popular in the recent literature (such as [10]).

Theorem 2 *Suppose $\mathcal{F}$ is a compact set in $C(\mathbf{X})$. There exists a strategy for Predictor that produces μ_n with $|\mu_n| \le 1$ and guarantees, for all $N = 1, 2, \dots$*

and all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \inf_{\epsilon \in (0,1/2]} \left(\mathcal{H}_\epsilon(\mathcal{F}) + \log\log \frac{1}{\epsilon} + \epsilon N + 1 \right), \tag{6}$$

where C is a universal constant.

Proof Without loss of generality we can only consider ϵ of the form 2^{-i}, $i = 1, 2, \ldots$, in (6). Let us fix, for each i, a 2^{-i}-net $\mathcal{F}_i$ for $\mathcal{F}$ of size $2^{\mathcal{H}_{2^{-i}}(\mathcal{F})}$; to each element of $\mathcal{F}_i$ we assign weight $\frac{6}{\pi^2} i^{-2} 2^{-\mathcal{H}_{2^{-i}}(\mathcal{F})}$, so that the weights sum to 1. Our goal (6) will be achieved if we establish, for each $i = 1, 2, \ldots$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \inf_{i=1,2,\ldots} \left(\mathcal{H}_{2^{-i}}(\mathcal{F}) + \log i + 2^{-i} N + 1 \right) \tag{7}$$

(we let C stand for different constants in different formulas).

Without loss of generality it will be assumed that F and all functions in $\mathcal{F}_i$, $i = 1, 2, \ldots$, take values in $[-1, 1]$. Fix an i. Let $F^* \in \mathcal{F}_i$ be 2^{-i}-close to F in $C(\mathbf{X})$. Lemma 1 gives a prediction strategy satisfying

$$\begin{aligned}
\sum_{n=1}^{N} (y_n - \mu_n)^2 &\le \sum_{n=1}^{N} (y_n - F^*(x_n))^2 + 8 \ln \left(\frac{\pi^2}{6} i^2 2^{\mathcal{H}_{2^{-i}}(\mathcal{F})} \right) \\
&\le \sum_{n=1}^{N} (y_n - F(x_n))^2 + 8 \ln \left(\frac{\pi^2}{6} i^2 2^{\mathcal{H}_{2^{-i}}(\mathcal{F})} \right) + 4 \left(2^{-i} N \right) \\
&\le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(1 + \log i + \mathcal{H}_{2^{-i}}(\mathcal{F}) + 2^{-i} N \right),
\end{aligned}$$

which coincides with (7). ∎

Banach function spaces as benchmark classes

Let $\mathcal{F}$ be a linear subspace of $C(\mathbf{X})$ equipped with a norm making it into a Banach space. We will be interested in the case where $\mathcal{F}$ is *compactly embedded* into $C(\mathbf{X})$, in the sense that the unit ball

$$U_{\mathcal{F}} := \{ F \in \mathcal{F} \mid \|F\|_{\mathcal{F}} \le 1 \}$$

is a compact subset of $C(\mathbf{X})$. (The Arzelà–Ascoli theorem, [17], 2.4.7, shows that all such $\mathcal{F}$ are Banach function spaces with finite embedding constant, as defined in [45] and below; in particular, they are proper Banach functional spaces.)

Theorem 3 *Let $\mathcal{F}$ be a Banach space compactly embedded in $C(\mathbf{X})$. There exists a strategy for Predictor that produces μ_n with $|\mu_n| \le 1$ and guarantees, for all $N = 1, 2, \ldots$ and all $F \in \mathcal{F}$,*

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \inf_{\epsilon \in (0,1/2]} \left(\mathcal{H}_{\epsilon/\phi}(U_{\mathcal{F}}) + \log\log\frac{1}{\epsilon} + \log\log\phi + \epsilon N + 1 \right), \quad (8)$$

where C is a universal constant and

$$\phi := 2\max\bigl(1, \|F\|_{\mathcal{F}}\bigr). \quad (9)$$

Proof Notice that $\mathcal{H}_\epsilon(2^i U_{\mathcal{F}}) = \mathcal{H}_{2^{-i}\epsilon}(U_{\mathcal{F}})$, $i = 1, 2, \ldots$. Applying (6) to $\mathcal{F} := 2^i U_{\mathcal{F}}$, we obtain

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(\mathcal{H}_{2^{-i}\epsilon}(U_{\mathcal{F}}) + \log\log\frac{1}{\epsilon} + \epsilon N + 1 \right) \quad (10)$$

for any $\epsilon \in (0, 1/2]$; we will assign weight $\frac{6}{\pi^2} i^{-2}$ to the corresponding prediction strategy. AA mixing the prediction strategies achieving (10) for $i = 1, 2, \ldots$, (it is clear that Lemma 1 is applicable to any prediction strategies, not only prediction rules), we obtain a strategy achieving

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(\mathcal{H}_{2^{-i}\epsilon}(U_{\mathcal{F}}) + \log\log\frac{1}{\epsilon} + \epsilon N + 1 \right) + 8\ln\left(\frac{\pi^2}{6} i^2\right) \quad (11)$$

for all $i = 1, 2, \ldots$ and all $F \in 2^i U_{\mathcal{F}}$. For each $F \in \mathcal{F}$ we can set

$$i := \max\left(1, \lceil \log \|F\|_{\mathcal{F}} \rceil\right)$$

to obtain $2^i \le \phi \le 2^{i+1}$ and so, from (11),

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(\mathcal{H}_{\epsilon/\phi}(U_{\mathcal{F}}) + \log\log\frac{1}{\epsilon} + \epsilon N + 1 \right) + 8\ln\left(\frac{\pi^2}{6} \log^2 \phi\right).$$

The last inequality can be written as (8). ∎

Competing with the continuous prediction rules

Let $\mathcal{F} \subseteq C(\mathbf{X})$ be a Banach function space (no connection between the norms in $\mathcal{F}$ and $C(\mathbf{X})$ is assumed) which is dense in $C(\mathbf{X})$ (in the $C(\mathbf{X})$ metric, of course); in this case we will say that $\mathcal{F}$ is *densely embedded* in $C(\mathbf{X})$. The *approachability* of $F \in C(\mathbf{X})$ by $\mathcal{F}$ is defined as the function

$$\mathcal{A}_\epsilon^{\mathcal{F}}(F) := \inf \left\{ \|F^*\|_{\mathcal{F}} \;\middle|\; \|F - F^*\|_{C(\mathbf{X})} \le \epsilon \right\}, \quad \epsilon > 0, \tag{12}$$

which is finite under our assumption of density.

Remark The *Gagliardo set* of a function $F \in C(\mathbf{X})$ can be defined as

$$\Gamma(F) := \Big\{ (t_0, t_1) \in \mathbb{R}^2 \;\Big|\; \exists F_0 \in C(\mathbf{X}), F_1 \in \mathcal{F} : F_0 + F_1 = F, \\ \|F_0\|_{C(\mathbf{X})} \le t_0, \|F_1\|_{\mathcal{F}} \le t_1 \Big\}. \tag{13}$$

(See [9], §3.1, for the general definition.) The graph of the function $\epsilon \mapsto \mathcal{A}_\epsilon^{\mathcal{F}}(F)$ is essentially the boundary of $\Gamma(F)$. A third way of talking about the Gagliardo set is in terms of the norm

$$K(t, F) := \inf_{F_0 \in \mathcal{F}, F_1 \in C(\mathbf{X}) : F = F_0 + F_1} \left(\|F_0\|_{C(\mathbf{X})} + t \|F_1\|_{\mathcal{F}} \right), \tag{14}$$

where t ranges over the positive numbers. (See [9], §3.1, or [2], 7.8, for further details.)

Theorem 4 *Let $\mathcal{F}$ be a Banach function space compactly and densely embedded in $C(\mathbf{X})$. Theorem 3's strategy guarantees, for all $N = 1, 2, \ldots$ and $F \in C(\mathbf{X})$,*

$$\sum_{n=1}^N (y_n - \mu_n)^2 \le \sum_{n=1}^N (y_n - F(x_n))^2 \\ + C \inf_{\epsilon \in (0, 1/2]} \left(\mathcal{H}_{\epsilon / A(\epsilon)}(U_{\mathcal{F}}) + \log\log \frac{1}{\epsilon} + \log\log A(\epsilon) + \epsilon N + 1 \right), \tag{15}$$

where C is a universal constant and $A(\epsilon) := 2 \max\left(1, \mathcal{A}_\epsilon^{\mathcal{F}}(F)\right)$.

Proof Inequality (8) immediately implies

$$\sum_{n=1}^N (y_n - \mu_n)^2 \le \sum_{n=1}^N (y_n - F(x_n))^2 \\ + C \inf_{\delta > 0} \inf_{\epsilon \in (0, 1/2]} \left(\mathcal{H}_{\epsilon / A(\delta)}(U_{\mathcal{F}}) + \log\log \frac{1}{\epsilon} + \log\log A(\delta) + \epsilon N + 4\delta N + 1 \right),$$

and it remains to restrict δ to $\delta \in (0, 1/2]$ and set $\epsilon := \delta$. ∎

Theorem 4 will be the source of many universal prediction strategies. Given any of the Banach spaces compactly and densely embedded in $C(\mathbf{X})$ introduced in §§5–6, Theorem 4 produces a universal prediction strategy: it is clear that (15) implies (1).

4 Finite-dimensional benchmark classes

We will be using (following [27]) the notation $f \sim g$ to mean $\lim_{\epsilon\to 0}(f(\epsilon)/g(\epsilon)) = 1$ and the notation $f \asymp g$ to mean $f = O(g)$ and $g = O(f)$ as $\epsilon \to 0$, where f and g are positive functions of $\epsilon > 0$.

If the benchmark class $\mathcal{F}$ is finite-dimensional, the typical rate of growth of its metric entropy is

$$\mathcal{H}_\epsilon(\mathcal{F}) \sim L \log \frac{1}{\epsilon}, \tag{16}$$

where L is the "metric dimension" of $\mathcal{F}$. This motivates the following corollaries of Theorems 2 and 3, respectively.

Corollary 1 *Suppose $\mathcal{F}$ is a compact set in $C(\mathbf{X})$ such that*

$$L := \limsup_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(\mathcal{F})}{\log \frac{1}{\epsilon}} \in (0, \infty). \tag{17}$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + CL \log N \tag{18}$$

from some N on, where C is a universal constant.

Proof It suffices to set $\epsilon := 1/N$ in (6). (And it is easy to check that this value of ϵ extracts from (6) an optimal, to within a constant factor, regret term.) ∎

Corollary 2 *Let $\mathcal{F}$ be a Banach space embedded in $C(\mathbf{X})$ and*

$$L := \limsup_{\epsilon\to 0} \frac{\mathcal{H}_\epsilon(U_\mathcal{F})}{\log \frac{1}{\epsilon}} \in (0, \infty). \tag{19}$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + CL \log N \tag{20}$$

from some N on, where C is a universal constant.

Remember that any Banach spaces $\mathcal{F}$ satisfying (19) is automatically finite-dimensional ([27], Theorem XII).

Proof of Corollary 2 Since the Banach space $\mathcal{F}$ is finite-dimensional, it is

compactly embedded in $C(\mathbf{X})$. Substituting $\epsilon := 1/N$ in (8), we obtain

$$\sum_{n=1}^{N} (y_n - \mu_n)^2$$
$$\leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C\left(L \log(\phi N) + \log\log N + \log\log \phi + 2\right)$$
$$\leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + 2CL \log N$$

from some N on. ∎

Theorem 4 is irrelevant to this section: no finite-dimensional subspace can be dense in $C(\mathbf{X})$ (since finite-dimensional subspaces are always closed).

Comparison with known results

It is instructive to compare the bound of Corollary 2 with a standard bound in competitive linear regression, obtained in [42] for the prediction strategy referred to as AAR in [42] and as the "Vovk–Azoury–Warmuth forecaster" in [12]. In the metric entropy method the elements of a net in $\mathcal{F}$ (the union of ϵ-nets of different balls in $\mathcal{F}$ for different ϵ, in the case of Theorem 3 and its corollaries) are AA mixed. AAR is conceptually very similar: instead of AA mixing the elements of the net, it AA mixes $\mathcal{F}$ itself; the weights assigned to the elements of the net are replaced by a "prior" probability measure on $\mathcal{F}$, and so summation is replaced by integration. An advantage of this "integration method" is that, for a suitable choice of the prior measure, it may produce a computationally efficient prediction strategy: e.g., AAR, which uses a Gaussian measure as prior, turned out to be a simple modification of ridge regression, as computationally efficient as ridge regression itself.

Suppose that $\mathbf{X}$ is a bounded subset of $\mathbb{R}^m$ and set

$$X_2 := \sup_{x \in \mathbf{X}} \|x\|_2, \quad X_\infty := \sup_{x \in \mathbf{X}} \|x\|_\infty; \tag{21}$$

it is clear that $X_2 \leq X_\infty$. AAR guarantees

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - \langle \theta, x_n \rangle)^2 + \|\theta\|_2^2 + m \ln\left(N X_\infty^2 + 1\right) \tag{22}$$

(see [42], (22) with $a := 1$ and Y^2 replaced by 1). To extract a similar inequality from (20), let U_m be the unit ball in $\mathbb{R}^m$ equipped with the $\|\cdot\|_2$ norm, $\mathcal{F}$ be the set of linear functions $x \in \mathbf{X} \mapsto \langle \theta, x \rangle$, $\theta \in \mathbb{R}^m$, with the norm $\|\theta\|_2$, and notice that

$$\mathcal{H}_\epsilon(U_\mathcal{F}) \leq X_2 \mathcal{H}_\epsilon(U_m) \leq X_2 \log \left\lceil \left(\frac{4}{\epsilon}\right)^m \right\rceil. \tag{23}$$

The first inequality in (23) follows from X_2 being the embedding constant of $\mathcal{F}$ into $C(\mathbf{X})$ (and also from the Cauchy–Schwarz inequality). The second inequality in (23) follows from the inequality (1.1.10) in [10].

Remark A popular alternative (used in [10] and, in a slightly modified form, [18]) to the notion of metric entropy $\mathcal{H}_\epsilon(A)$ is that of *entropy numbers* $\epsilon_n(A)$, $n = 1, 2, \ldots$, defined as the infimum of ϵ such that there exists an ϵ-net for A. Notice that the "infimum" here is attained (and so can be replaced by "minimum") because of the compactness of A^n. It is easy to see that

$$2^{\mathcal{H}_\epsilon(A)} = \min\{n \,|\, \epsilon_n(A) \le \epsilon\}; \tag{24}$$

this can be useful when translating results about entropy numbers into results about metric entropy.

Combining (23) with Corollary 2, we obtain the following analogue of (22).

Corollary 3 *Let* $\mathbf{X}$ *be a bounded set in* $\mathbb{R}^m$ *and* X_2 *be defined by (21). There exists a strategy for Predictor that guarantees, for all* $\theta \in \mathbb{R}^m$*,*

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - \langle \theta, x_n \rangle)^2 + C X_2 m \log N \tag{25}$$

from some N *on, where* C *is a universal constant.*

An interesting feature of the regret terms in (22) and (25) is their logarithmic dependence on N; some other standard bounds, such as those in [11], [24] and [6], involve $\sqrt{N}$ (or similar terms, such as the square root of the competitor's loss). It is remarkable that the bound established in the first paper on competitive on-line regression, [20], also depends on N logarithmically; the method used in that paper is penalized minimum least squares. An important advantage of the bounds given in [11, 24, 6] is that the character of their dependence on the dimension m allows one to carry them over to infinite-dimensional function spaces; these bounds will be discussed again in §6.

5 Benchmark classes of analytic functions

In this section we consider classes of analytic functions, and so it is natural to consider complex-valued functions of one or more complex variables. The observations are now any complex numbers, $y_n \in \mathbb{C}$, bounded by 1 in absolute value, and so prediction rules are functions $F : \mathbf{X} \to \mathbb{C}$. Also, in this section $C(\mathbf{X})$ will stand for the function space of continuous complex-valued functions on $\mathbf{X}$. It is clear that Theorems 1–4 continue to hold in this extended framework.

According to [27], §3.II, the typical growth rate for the metric entropy of infinite-dimensional classes $\mathcal{F}$ of analytic functions on $\mathbf{X}$ is

$$\mathcal{H}_\epsilon(\mathcal{F}) \asymp \log^{m+1} \frac{1}{\epsilon}, \tag{26}$$

where m is the dimension of $\mathbf{X}$. (Although intermediate rates such as

$$\mathcal{H}_\epsilon(\mathcal{F}) \asymp \frac{\log^{m+1} \frac{1}{\epsilon}}{\log\log \frac{1}{\epsilon}}$$

also sometimes occur.) For such growth rates (the complex versions of) Theorems 2–4 imply the following three corollaries.

Corollary 4 *Suppose $\mathcal{F}$ is a compact set in $C(\mathbf{X})$ and $M > 0$ is such that*

$$L := \limsup_{\epsilon \to 0} \frac{\mathcal{H}_\epsilon(\mathcal{F})}{\log^M \frac{1}{\epsilon}} \in (0, \infty).$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + CL \log^M N \tag{27}$$

from some N on, where C is a universal constant.

Proof As in the proof of Corollary 1, set $\epsilon := 1/N$ in (6). ∎

Corollary 5 *Let $\mathcal{F}$ be a Banach function space compactly embedded in $C(\mathbf{X})$ and $M > 0$ be a number such that*

$$L := \limsup_{\epsilon \to 0} \frac{\mathcal{H}_\epsilon(U_\mathcal{F})}{\log^M \frac{1}{\epsilon}} \in (0, \infty). \tag{28}$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + CL \log^M N \tag{29}$$

from some N on, where C is a universal constant.

Proof Following the proof of Corollary 2, we substitute $\epsilon := 1/N$ in (8) to obtain, from some N on:

$$\begin{aligned}
&\sum_{n=1}^{N} |y_n - \mu_n|^2 \\
&\quad \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C\left(L \log^M(\phi N) + \log\log N + \log\log\phi + 2\right) \\
&\quad \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C' L \log^M N,
\end{aligned}$$

where C' is another universal constant. ∎

Unlike in the previous section, Theorem 4 is not vacuous for classes of analytic functions: as we will see in the following subsection, there are numerous examples of such classes that are compactly and densely embedded in $C(\mathbf{X})$, for important signal spaces $\mathbf{X}$. The following is the implication of Theorem 4 for the growth rate (26); unfortunately, this statement still has $\inf_\epsilon$ since the growth rate of $\mathcal{A}^{\mathcal{F}}_\epsilon(F)$ is unknown.

Corollary 6 *Let $\mathcal{F}$ be a Banach function space compactly and densely embedded in $C(\mathbf{X})$ and let L and M be positive numbers satisfying (28). There exists a strategy for Predictor that guarantees, for all $F \in C(\mathbf{X})$,*

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C_M \inf_{\epsilon \in (0,1]} \left(L \left(\log^+ \mathcal{A}^{\mathcal{F}}_\epsilon(F) \right)^M + L \log^M \frac{1}{\epsilon} + \epsilon N \right) \quad (30)$$

from some N on, where C_M is a constant depending only on M and $\log^+$ is defined as

$$\log^+ t := \begin{cases} \log t & \text{if } t \ge 1 \\ 0 & \text{otherwise.} \end{cases}$$

Proof It is clear that the optimal value of ϵ in the regret term in (15) tends to 0 as $N \to \infty$, and so the regret term can be bounded above by

$$C \inf_{\epsilon \in (0,1/2]} \left(L \log^M \left(\frac{A(\epsilon)}{\epsilon} \right) + \log\log \frac{1}{\epsilon} + \log\log A(\epsilon) + \epsilon N + 1 \right) \le C' \inf_{\epsilon \in (0,1]} \left(L \left(\log^+ \mathcal{A}^{\mathcal{F}}_\epsilon(F) + \log \frac{1}{\epsilon} \right)^M + \epsilon N \right)$$

from some N on. (The case $F \in \mathcal{F}$ has to be considered separately.) ∎

Examples

We will reproduce two simple examples from [27]; for simplicity we only consider analytic functions of one complex variable (although already [27] contains results making extension to several variables straightforward). Remember that the set of all complex numbers is denoted $\mathbb{C}$.

Let K be a simply connected continuum in $\mathbb{C}$ containing more than one point and G be a region (connected open set) such that $K \subseteq G \subseteq \mathbb{C}$. The set of all complex-valued functions on K that admit a bounded analytic continuation to G is denoted A^K_G. Equipped with the usual pointwise addition and scalar action and with the norm

$$\|f|_K\|_{A^K_G} := \sup_{z \in G} |f(z)|, \quad (31)$$

where $f : G \to \mathbb{C}$ ranges over the bounded analytic functions and $f|_K$ is the restriction of f to K, it becomes a Banach space. Expression (31) is well-defined by the uniqueness theorem in complex analysis, and the completeness of A_G^K follows from the fact (known as Weierstrass's theorem, [3], Theorem IV.1.1) that uniform limits of analytic functions are analytic.

It is shown in [27], (139), that

$$\mathcal{H}_\epsilon \left(U_{A_G^K} \right) \sim \tau(G,K) \log^2 \frac{1}{\epsilon} \tag{32}$$

(this was hypothesised by Kolmogorov and proved independently by Babenko and Erokhin; in [51] the constant $\tau(G,K)$ was shown, under mild restrictions, to be proportional to the Green capacity of K relative to G). Therefore, Corollary 5 gives a strategy for Predictor guaranteeing

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C\tau(G,K) \log^2 N \tag{33}$$

for all $F \in A_G^K$ and from some N on, where C is a universal constant.

In many interesting special cases considered in [27], §7, the constant $\tau(G,K)$ has a simple explicit expression, e.g.:

- $\tau(G,K) = 1/\log(R/r)$ if $K = r\overline{\mathbb{D}}$ and $G = R\mathbb{D}$, $R > r > 0$, $\mathbb{D} := U_{\mathbb{C}}$ being the open unit disk in $\mathbb{C}$;
- $\tau(G,K) = 1/(2\log\lambda)$ if $K = [-1,1]$ and G is the ellipse with the sum of semi-axes equal to $\lambda > 1$ and with foci at the points ± 1 (there is a misprint in [27], (131); the correct formula is given in, e.g., [41], Theorem 1 in §12).

Both these expressions were obtained by Vitushkin.

Let $h > 0$. The vector space of all periodic period 2π complex-valued functions on the real line $\mathbb{R}$ that admit a bounded analytic continuation to the strip $\{z \in \mathbb{C} \,|\, |\operatorname{Im} z| < h\}$ is denoted A_h. The norm in this space is defined by

$$\|f|\mathbb{R}\|_{A_h} := \sup_{z:|\operatorname{Im} z|<h} |f(z)| , \tag{34}$$

where f ranges over the bounded analytic functions on $\{z \,|\, |\operatorname{Im} z| < h\}$. Expression (34) is again well-defined and the normed space A_h is complete. The estimate of the metric entropy of the unit ball of A_h given in [27], (130), is

$$\mathcal{H}_\epsilon \left(U_{A_h} \right) \sim \frac{2}{h \log e} \log^2 \frac{1}{\epsilon} \tag{35}$$

(Vitushkin). Corollary 5 now gives

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + \frac{C}{h} \log^2 N \tag{36}$$

for all $F \in A_h$ and from some N on, where C is a universal constant.

Corollary 6 is applicable to $\mathcal{F} = A_h$ for any $h > 0$, and so A_h gives rise to a universal prediction strategy. Indeed, taking $\mathbf{X} := \partial\mathbb{D}$ (the unit circle in $\mathbb{C}$) and identifying complex-valued functions on $\partial\mathbb{D}$ with the corresponding periodic period 2π complex-valued functions on $\mathbb{R}$ (namely, $f : \partial\mathbb{D} \to \mathbb{C}$ is identified with the function $t \in \mathbb{R} \mapsto f\left(e^{it}\right)$), we can arbitrarily closely in $C(\mathbf{X})$ approximate each $F \in C(\mathbf{X})$ by a trigonometric polynomial (this is Weierstrass's second theorem, [1], §21), whose analytic continuation to $\{z \,|\, |\operatorname{Im} z| < h\}$ is bounded; we can see that A_h is dense in $C(\mathbf{X})$.

Suppose $\mathbf{X} \subseteq \mathbb{C}$ is compact (in particular, closed). For $A_G^{\mathbf{X}}$ to be dense in $C(\mathbf{X})$, $\mathbf{X}$ must be nowhere dense in $\mathbb{C}$ (since limits in $C(\mathbf{X})$ of elements of $A_G^{\mathbf{X}}$ would be analytic in the interior points of $\mathbf{X}$). If we additionally assume that $\mathbf{X}$ is simply connected, Mergelyan's theorem ([32], Theorem 20.5) will guarantee that every continuous complex-valued function on $\mathbf{X}$ can be arbitrarily closely in $C(\mathbf{X})$ approximated by a polynomial. We can see that $A_G^{\mathbf{X}}$ is dense in $C(\mathbf{X})$ provided $\mathbf{X}$ is a nowhere dense simply connected compact. The most interesting case is perhaps where $\mathbf{X} = [a, b]$ is a closed interval in $\mathbb{R}$.

Dense function spaces popular in learning theory

Benchmark classes such as A_G^K and A_h have never been used, to my knowledge, in competitive on-line prediction. Familiar rates of growth of the regret term are $O(\log N)$ or N^α (for $\alpha \in (0, 1)$, usually $\alpha = 1/2$); intermediate rates obtainable for A_G^K and A_h, such as (33) and (36), have not been known.

Several benchmark classes of this type, however, have been implicitly considered since they are reproducing kernel Hilbert spaces corresponding to popular reproducing kernels (see [40] and [35] for the use of reproducing kernels in learning theory and [5] for the theory of reproducing kernel Hilbert spaces, or RKHS for brevity). One of such spaces is the Hardy space $H^2(\mathbb{D})$ restricted to the interval $(-1, 1)$ of the real line (see, e.g., [30]). Mergelyan's theorem (or Weierstrass's first theorem, [1], §20) immediately implies that for each $\epsilon > 0$ the restriction of $H^2(\mathbb{D})$ to $[-1 + \epsilon, 1 - \epsilon]$ is dense in $C([-1 + \epsilon, 1 - \epsilon])$: indeed, each polynomial belongs to $H^2(\mathbb{D})$. (In the multi-dimensional case, this fact was established by Steinwart [36], Example 2.) It is easy to see that, when $\mathbf{X} = [-1 + \epsilon, 1 - \epsilon]$, (33) holds not only for $A_G^K := A_{\mathbb{D}}^{\mathbf{X}}$ but also for A_G^K replaced by the restriction of $H^2(\mathbb{D})$ to $\mathbf{X}$ and for $\tau(G, K)$ replaced by a suitable constant depending only on ϵ.

Remark The reproducing kernel

$$\mathbf{K}(z, w) := \frac{1}{1 - \overline{w}z}$$

of the Hardy space $H^2(\mathbb{D})$ is known as the Szegö kernel. In some recent learning literature (such as [36], Example 2, [35], Example 4.24) the restriction of the multidimensional Szegö kernel to the unit ball in a Euclidean space is referred to as "Vovk's infinite-degree polynomial kernel". The origin of this undeserved

name is the SVM manual [34]; I liked to use the Szegö kernel when explaining the idea of reproducing kernels to my students.

Other popular spaces of analytic functions on $\mathbb{R}^m$ are the reproducing kernel Hilbert spaces corresponding to the "Gaussian RBF kernels", parameterized by $\sigma > 0$. They are described in [37] (and also earlier in [7] and, more explicitly, [33]). We have for them both the $O(\log^2 N)$ rate of growth of the regret term and the denseness in $C(K)$ for each compact $K \subseteq \mathbb{R}^m$ (see [36], Example 1). Interestingly, these RKHS do not look dense in $C(K)$: it appears that they can only approximate functions at the scale comparable with the parameter σ (perhaps the cause of this illusion is the small metric entropy of these function classes).

In general, it appears that most common reproducing kernels give rise to RKHS consisting of analytic functions. Suppose that $\mathbf{X}$ is a bounded set in a Euclidean space $\mathbb{R}^m$. It is often the case that the reproducing kernel $\mathbf{K}(z, w)$, $z, w \in \mathbf{X}$, admits a continuation to a neighbourhood $O^2 \subseteq (\mathbb{C}^m)^2$ of $\overline{\mathbf{X}}^2$ analytic in its first argument z and remaining a reproducing kernel. By the Tietze–Uryson theorem ([19], 2.1.8) there is an intermediate neighbourhood G, such that $\overline{\mathbf{X}} \subseteq G \subseteq \overline{G} \subseteq O$. Let $\mathcal{F}$ be the RKHS on G generated by the given reproducing kernel $\mathbf{K}$ thus extended to G^2. It is clear that

$$\mathbf{c}_{\mathcal{F}} := \sup_{z \in G} \sqrt{\mathbf{K}(z, z)} \tag{37}$$

is finite. The set of the evaluation functionals $\mathbf{K}_w(z) := \mathbf{K}(z, w)$, $w \in G$, is dense in $\mathcal{F}$ ([5], §2(4); for details, see [4], Theorem 2), convergence in $\mathcal{F}$ implies convergence in $C(G)$ (by $\mathbf{c}_{\mathcal{F}} < \infty$ and [5], §2(5)), each $\mathbf{K}_w$ is analytic, and uniform limits of analytic functions are analytic ([3], Theorem IV.1.1); therefore, $\mathcal{F}$ consists of analytic functions. Since $\mathbf{c}_{\mathcal{F}} < \infty$, we have $U_{\mathcal{F}} \subseteq \mathbf{c}_{\mathcal{F}} U_{A_G^{\mathbf{X}}}$, and so $\mathcal{F}$ is compactly embedded in $C(\mathbf{X})$ and, as above, the regret term grows as a polynomial of $\log N$.

Steinwart [36] gives four examples of reproducing kernels on $\mathbf{X}^2$ that can be analytically continued to a neighbourhood of $\mathbf{X}^2$, as in the previous paragraph, and whose RKHS are dense in $C(\mathbf{X})$ (we described the first two of his examples above).

Sometimes formulas for reproducing kernels contain "awkward" building blocks such as taking the fractional part ([50], (10.2.4)), absolute value, or min ([43], (8)), and in this case analytic continuation to a neighbourhood is usually impossible. Such reproducing kernels are often derived from the corresponding RKHS that are much more massive than the classes of analytic functions considered in this section; such massive classes will be considered in the next section.

6 Sobolev-type classes

We now return to our basic prediction protocol in which the observations y_n are real numbers (bounded by 1 in absolute value); $C(\mathbf{X})$ will again denote the

continuous real-valued functions on $\mathbf{X}$.

Typical classes studied in the theory of functions of real variable are much more massive than typical classes of analytic functions. In the second part of this section we will see examples showing that the typical growth rate for the metric entropy of compact classes $\mathcal{F}$ of real-valued functions defined on nice subsets of a Euclidean space is

$$\mathcal{H}_\epsilon(\mathcal{F}) \asymp (1/\epsilon)^\gamma, \tag{38}$$

where $\gamma > 0$ is the "degree of non-smoothness" of $\mathcal{F}$. The following two corollaries are asymptotic versions of Theorems 2–3 for this growth rate.

Corollary 7 *Suppose a compact set $\mathcal{F}$ in $C(\mathbf{X})$ and a positive number γ satisfy*

$$L := \limsup_{\epsilon \to 0} \frac{\mathcal{H}_\epsilon(\mathcal{F})}{(1/\epsilon)^\gamma} \in (0, \infty).$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^N (y_n - \mu_n)^2 \le \sum_{n=1}^N (y_n - F(x_n))^2 + CL^{\frac{1}{\gamma+1}} N^{\frac{\gamma}{\gamma+1}} \tag{39}$$

from some N on, where C is a universal constant.

Proof Solving

$$L(1/\epsilon)^\gamma + \epsilon N \to \min, \tag{40}$$

we obtain

$$\epsilon = \left(\frac{L\gamma}{N}\right)^{\frac{1}{\gamma+1}} \to 0 \quad (N \to \infty) \tag{41}$$

and

$$L(1/\epsilon)^\gamma + \epsilon N = \left(\gamma^{\frac{1}{\gamma+1}} + \gamma^{-\frac{\gamma}{\gamma+1}}\right) L^{\frac{1}{\gamma+1}} N^{\frac{\gamma}{\gamma+1}}; \tag{42}$$

since the first factor on the right-hand side of the last expression always belongs to $(1, 2]$, it can be ignored. ∎

Remark We will have to find minima such as (40) on several occasions, and for the future reference I will give the general result of the calculation for

$$A\epsilon^{-a} + B\epsilon^b \to \min, \tag{43}$$

where A, B, a, b are positive numbers and ϵ ranges over $(0, \infty)$. The minimum is attained at

$$\epsilon = \left(\frac{Aa}{Bb}\right)^{\frac{1}{a+b}} \tag{44}$$

and is equal to

$$\left((a/b)^{\frac{b}{a+b}} + (b/a)^{\frac{a}{a+b}}\right) A^{\frac{b}{a+b}} B^{\frac{a}{a+b}}. \tag{45}$$

Instead of finding the precise minimum in (43), it will usually be more convenient to approximate it by equating the two addends in (43), which gives

$$\epsilon = (A/B)^{\frac{1}{a+b}} \tag{46}$$

and so gives the upper bound

$$2A^{\frac{b}{a+b}}B^{\frac{a}{a+b}} \tag{47}$$

for (45).

Corollary 8 *Let $\mathcal{F}$ be a Banach function space compactly embedded in $C(\mathbf{X})$ and γ be a positive number such that*

$$L := \limsup_{\epsilon \to 0} \frac{\mathcal{H}_\epsilon(U_\mathcal{F})}{(1/\epsilon)^\gamma} \in (0, \infty). \tag{48}$$

There exists a strategy for Predictor that guarantees, for all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + CL^{\frac{1}{\gamma+1}}\phi^{\frac{\gamma}{\gamma+1}}N^{\frac{\gamma}{\gamma+1}} \tag{49}$$

from some N on, where C is a universal constant and ϕ is defined by (9).

Proof Substituting $L\phi^\gamma$ for L on the right-hand side of (42) and ignoring the first factor on the right-hand side, we obtain

$$(L\phi^\gamma)^{\frac{1}{\gamma+1}}N^{\frac{\gamma}{\gamma+1}} = L^{\frac{1}{\gamma+1}}\phi^{\frac{\gamma}{\gamma+1}}N^{\frac{\gamma}{\gamma+1}}. \qquad \blacksquare$$

Examples

We will say that a function F defined on a metric space with metric ρ is *Hölder continuous of order $\alpha \in (0, 1]$ with coefficient $c > 0$* if, for all x and x' in the domain of F, $|F(x) - F(x')| \le c\rho^\alpha(x, x')$. If $\alpha = 1$, we will also say that F is *Lipschitzian with coefficient c*.

Let $\mathbf{X}$ be an m-dimensional (axes-parallel) parallelepiped. Define $\mathcal{F}$ to be the class of real-valued functions on $\mathbf{X}$ that are bounded in $C(\mathbf{X})$ by a given constant and whose kth partial derivatives exist and are all Hölder continuous of order α with a given coefficient. It is shown in [27], Theorem XIII, that

$$\mathcal{H}_\epsilon(\mathcal{F}) \asymp (1/\epsilon)^\gamma, \tag{38}$$

where $\gamma := m/s = m/(k + \alpha)$ is the "degree of non-smoothness" ($1/\gamma$ was called the "degree of smoothness" by G. G. Lorentz in his review of [27] in *Mathematical Reviews*) and $s := k + \alpha$ is the "indicator of smoothness" ([27], §3.III). We can now deduce from (39) that

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_\mathcal{F} N^{\frac{m}{m+s}} \tag{50}$$

for all $F \in \mathcal{F}$ and from some N on, where $C_{\mathcal{F}}$ is a constant depending on $\mathcal{F}$ but nothing else.

For the class $\mathcal{F}$ of Lipschitzian functions with coefficient c defined on an interval of the real line of length l and bounded in absolute value by a given constant Kolmogorov and Tikhomirov [27] (see their (10), which also remains true when $\mathcal{H}_\epsilon(A)$ is replaced by $\mathcal{H}_\epsilon^A(A)$) obtain the more accurate estimate $\mathcal{H}_\epsilon(\mathcal{F}) \sim cl/\epsilon$. In this case (50) can be replaced by

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C\sqrt{cl}N^{1/2}, \tag{51}$$

where C is a universal constant.

Results of this type have been greatly extended in recent years. We will later state one such result about Besov spaces $B^s_{p,q}(\mathbf{X})$. For the general definition of Besov spaces see [18], §§2.2,2.5. Besov spaces $B^s_{p,q}$ are Banach spaces (assuming $p, q \geq 1$), but we will consider them as topological vector spaces (i.e., will regard Banach spaces with equivalent norms as the same space).

Remark A popular definition of Besov spaces is via "real interpolation" (as in [2], Chapter 7). For example, according to this definition, $B^s_{p,\infty}(\mathbf{X})$, where $s \in (0, \infty)$ and $p \in [1, \infty)$, consists of the functions F whose Gagliardo set (13) with $C(\mathbf{X})$ replaced by $L^p(\mathbf{X})$ and $\mathcal{F}$ replaced by the Sobolev space $W^{m,p}(\mathbf{X})$ (see [2], Chapter 3) for some integer $m > s$ contains the curve

$$\left\{(t_0, t_1) \in \mathbb{R}^2 \;\middle|\; t_0^{1-\theta} t_1^{\theta} = c\right\}, \quad \theta := s/m,$$

for some positive c; the infimum of c with this property is the norm of F in $B^s_{p,\infty}(\mathbf{X})$.

We are only interested in Besov spaces whose domain is the signal space $\mathbf{X}$. In the rest of this section it will always be assumed that $\mathbf{X}$ is a subset of Euclidean space, $\mathbf{X} \subseteq \mathbb{R}^m$, which is a *minimally regular domain*, in the sense that it is bounded and coincides with the interior of its closure ([18], Definition 2.5.1/2).

Every $B^s_{p,q}(\mathbf{X})$ with $s > m/p$ is compactly embedded in $C(\mathbf{X})$ (apply [18], (2.5.1/10), to $s_1 := s$, $p_1 := p$, $q_1 := q$, $p_2 := q_2 := \infty$ and sufficiently small $s_2 > 0$ and remember that $\mathscr{C}^s(\mathbf{X}) := B^s_{\infty,\infty}(\mathbf{X})$ are Hölder–Zygmund spaces, [18], 2.2.2(iv)). We will be interested only in this case. Edmunds and Triebel's general result (Theorem 3.5 of [18] applied to $s_1 := s$, $p_1 := p$, $q_1 := q$, $s_2 := 0$, $p_2 := \infty$ and $q_2 := 1$, in combination with (2.3.3/3)) then shows that

$$\mathcal{H}_\epsilon\left(U_{B^s_{p,q}(\mathbf{X})}\right) \asymp (1/\epsilon)^{m/s}$$

(use (24) to move between entropy numbers and metric entropy). We can see that (50) still holds for $\mathcal{F}$ a bounded set in a general Besov space $B^s_{p,q}(\mathbf{X})$;

moreover, Corollary 8 shows that

$$\sum_{n=1}^{N}(y_n-\mu_n)^2 \le \sum_{n=1}^{N}(y_n-F(x_n))^2 + C_{\mathbf{X},s,p,q}\left(\|F\|_{B^s_{p,q}(\mathbf{X})}+1\right)^{\frac{m}{m+s}} N^{\frac{m}{m+s}} \tag{52}$$

for all $F \in B^s_{p,q}(\mathbf{X})$ from some N on, where $C_{\mathbf{X},s,p,q}$ is a constant depending only on $\mathbf{X}, s, p, q$. Setting p and q to ∞, we recover (50).

To conclude this subsection, let us go back to reproducing kernels. Cucker and Smale ([16], Theorem D) show that if $\mathcal{F}$ is an RKHS with a C^∞ reproducing kernel on $\mathbf{X}^2$ for a compact set $\mathbf{X} \subseteq \mathbb{R}^m$,

$$\mathcal{H}_\epsilon(U_\mathcal{F}) = O\left((1/\epsilon)^{2m/h}\right)$$

for an arbitrary $h > m$. Corollary 8 (together with its proof, since the L in (48) is 0 for each γ and so has to be replaced by an upper bound) shows that, for an arbitrarily small $\delta > 0$,

$$\sum_{n=1}^{N}(y_n-\mu_n)^2 \le \sum_{n=1}^{N}(y_n-F(x_n))^2 + N^\delta \tag{53}$$

for all $F \in \mathcal{F}$ from some N on. The regret term in (53) is not as good as the poly-log regret term for RKHS with analytic reproducing kernels (see p. 17), but this is not surprising: the class of analytic functions is known to be much narrower than that of infinitely differentiable functions (for a useful relation between these classes see [38], 3.7.1).

Comparisons with defensive forecasting

Many of the Besov spaces $B^s_{p,q}(\mathbf{X})$ are "uniformly convex", and this makes it possible to apply to them a result obtained in [45] using the method of "defensive forecasting".

Let V be a Banach space and $\partial U_V := \{v \in V \mid \|v\|_V = 1\}$ be the unit sphere in V (the boundary of the unit ball U_V). A convenient measure of rotundity of the unit ball U_V is Clarkson's [13] modulus of convexity

$$\delta_U(\epsilon) := \inf_{\substack{u,v\in\partial U_V \\ \|u-v\|_V=\epsilon}} \left(1-\left\|\frac{u+v}{2}\right\|_V\right), \quad \epsilon \in (0,2] \tag{54}$$

(we will be mostly interested in the small values of ϵ).

If a Banach space $\mathcal{F}$ is continuously embedded in $C(\mathbf{X})$, the embedding constant will be denoted $\mathbf{c}_\mathcal{F}$:

$$\mathbf{c}_\mathcal{F} := \sup_{F\in U_\mathcal{F}} \|F\|_{C(\mathbf{X})} < \infty \tag{55}$$

(we have already used this notation in the special case of RKHS: cf. (37)).

Proposition 1 ([45], Theorem 1) *Let $\mathcal{F}$ be a Banach space continuously embedded in $C(\mathbf{X})$ and such that*

$$\forall \epsilon \in (0,2] : \delta_{\mathcal{F}}(\epsilon) \geq (\epsilon/2)^p/p \tag{56}$$

for some $p \in [2,\infty)$. There exists a strategy for Predictor producing μ_n that are guaranteed to satisfy

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + 40\sqrt{\mathbf{c}_{\mathcal{F}}^2 + 1}\,(\|F\|_{\mathcal{F}} + 1)\, N^{1-1/p} \tag{57}$$

for all $N = 1, 2, \ldots$ and all $F \in \mathcal{F}$.

It is interesting that in Proposition 1 $\mathcal{F}$ is not required to be compactly embedded in $C(\mathbf{X})$.

It was shown by Clarkson ([13], §3) that, for $p \in [2,\infty)$,

$$\delta_{L^p}(\epsilon) \geq 1 - \left(1 - (\epsilon/2)^p\right)^{1/p}.$$

(And this bound was shown to be optimal in [22].) This result was extended to some other Besov spaces in [15], Theorem 3: the modulus of convexity of each Besov space $B^s_{p,q}(\mathbb{R}^m)$, $s \in \mathbb{R}$, $p \in [2,\infty)$ and $q \in [p/(p-1), p]$, also satisfies

$$\delta_{B^s_{p,q}(\mathbb{R}^m)}(\epsilon) \geq 1 - \left(1 - (\epsilon/2)^p\right)^{1/p}. \tag{58}$$

Edmunds and Triebel [18], 2.5.1, define the Besov space $B^s_{p,q}(\mathbf{X})$ on $\mathbf{X} \subseteq \mathbb{R}^m$ as the set of all restrictions of the functions in $B^s_{p,q}(\mathbb{R}^m)$ to $\mathbf{X}$ with the norm

$$\|F\|_{B^s_{p,q}(\mathbf{X})} := \inf_{F^*} \|F^*\|_{B^s_{p,q}(\mathbb{R}^m)}, \tag{59}$$

where F^* ranges over all extensions of F to $\mathbb{R}^m$. To check that $B^s_{p,q}(\mathbf{X})$ is at least as convex as $B^s_{p,q}(\mathbb{R}^m)$ for $p \geq 2$ and $p/(p-1) \leq q \leq p$, take any $F_1, F_2 \in B^s_{p,q}(\mathbf{X})$ of norm 1 and at a distance of ϵ from each other. If the infima in the definition (59) of the norms of F_1 and F_2 are attained, we can take the extensions F_1^* and F_2^* to $\mathbb{R}^m$ of norm 1 and notice that, as $\|F_1^* - F_2^*\|_{B^s_{p,q}(\mathbb{R}^m)} \geq \epsilon$ and the modulus of convexity is a non-decreasing function of ϵ ([28], Lemma 1.e.8),

$$\left\|\frac{F_1 + F_2}{2}\right\|_{B^s_{p,q}(\mathbf{X})} \leq \left\|\frac{F_1^* + F_2^*}{2}\right\|_{B^s_{p,q}(\mathbb{R}^m)} \leq 1 - \delta_{B^s_{p,q}(\mathbb{R}^m)}(\epsilon).$$

If the infima are not attained, we can still use a similar argument for $p \geq 2$ and $p/(p-1) \leq q \leq p$ with $\delta_{B^s_{p,q}(\mathbb{R}^m)}$ replaced by its lower bound given by (58). This shows that (58) extends to arbitrary domains:

$$\delta_{B^s_{p,q}(\mathbf{X})}(\epsilon) \geq 1 - \left(1 - (\epsilon/2)^p\right)^{1/p} \geq (\epsilon/2)^p/p. \tag{60}$$

Let $p \in [2, \infty)$, $q \in [p/(p-1), p]$ and $s \in (m/p, \infty)$. By Proposition 1 and (60), there exist a constant $C_{\mathbf{X},s,p,q} > 0$ and a strategy for Predictor producing μ_n that are guaranteed to satisfy

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{\mathbf{X},s,p,q} \left(\|F\|_{B^s_{p,q}} + 1 \right) N^{1-1/p} \tag{61}$$

for all $N = 1, 2, \ldots$ and all $F \in B^s_{p,q}(\mathbf{X})$. We can see that defensive forecasting works better than metric entropy at the "wild" end of the scale $B^s_{p,q}(\mathbf{X})$ whereas metric entropy better copes with smooth functions (at this time we only pay attention to the exponent of N, which is more important, from the asymptotic point of view as $N \to \infty$, than the coefficient in front of $N^{\cdots}$):

- Suppose $s \in (m/p, m/2]$. The exponent $1 - 1/p$ of N in (61) can be taken arbitrarily close to $1 - s/m$, and we can see that it is then better than the exponent of N in (52):
$$1 - \frac{s}{m} < \frac{m}{m+s}.$$
For example, in the very important case $m = 1, s \approx 1/2$ (typical trajectories of the Brownian motion are of this type) defensive forecasting gives approximately $N^{1/2}$ whereas the method of metric entropy gives approximately $N^{2/3}$.

- Suppose $s \in (m/2, m)$. The exponent of N in (61) can always be taken as $1/2$, and it is still better than the exponent of N in (52):
$$\frac{1}{2} < \frac{m}{m+s}.$$

- Suppose $s \in [m, \infty)$. A weakness of the method of defensive forecasting (in its current state: see, e.g., [44] and [45], in addition to (61)) is that it cannot give regret terms better than $O(N^{1/2})$. Therefore, the method of metric entropy beats defensive forecasting for smooth Besov spaces $B^s_{p,q}(\mathbf{X})$, $s > m$.

For comparison with (50), define the norm

$$\|F\|_s := \max \left(\sup_{x \in \mathbf{X}} |F(x)|, \max_{|\beta|=k} \sup_{x, x' \in \mathbf{X} : x \neq x'} \frac{\left| D^\beta F(x) - D^\beta F(x') \right|}{\|x - x'\|^\alpha} \right), \tag{62}$$

where $\mathbf{X}$ is a parallelepiped in $\mathbb{R}^m$, $\beta = (\beta_1, \ldots, \beta_m)$ ranges over the multi-indices, $\|\cdot\|$ is any standard norm in $\mathbb{R}^m$, $F : \mathbf{X} \to \mathbb{R}$ is k times continuously differentiable function, $\alpha \in (0, 1]$, and $s := k + \alpha$. It is easy to check that the Banach space normed by (62) is continuously embedded in $B^{s'}_{p,2}(\mathbf{X})$ for any $s' < s$: indeed, it is obvious that the space normed by (62) is continuously embedded in the Hölder–Zygmund space $\mathscr{C}^s(\mathbf{X}) := B^s_{\infty,\infty}(\mathbf{X})$ ([18], 2.2.2(iv), [39], 1.2.2), and the usual embedding theorem implies that the Hölder–Zygmund

space is continuously embedded in $B^{s'}_{p,2}(\mathbf{X})$ ([18], (2.5.1/10), with $p_1 = q_1 = \infty$). Fixing an arbitrarily small $\delta > 0$, we deduce from (61) that for each $s \le m/2$ there exists a constant $C_{\mathbf{X},s,\delta} > 0$ such that

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{\mathbf{X},s,\delta} \left(\|F\|_s + 1 \right) N^{1-s/m+\delta} \tag{63}$$

for all $N = 1, 2, \ldots$ and all F with finite $\|F\|_s$. For s above $m/2$ we have to take $p = 2$ and so obtain

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C_{\mathbf{X},s} \left(\|F\|_s + 1 \right) N^{1/2} \tag{64}$$

in place of (63); (64) starts losing to (50) when s exceeds m.

It is also interesting to compare (64) for $m = s = 1$ with (51). Even though $\|F\|_{C(\mathbf{X})} \le 1$ is the only interesting case, the bound in (51) still appears better: it scales as $\sqrt{c}$ in c, whereas (64) scales as c when applied to $\{F \mid \|F\|_s \le c\}$. This impression is confirmed by a more careful analysis: (49) implies

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \sqrt{l \left(\|F\|_s + 1 \right) N}$$

for all F with finite $\|F\|_s$ from some N on, where C is a universal constant. Comparing this with (64), we can see another disadvantage of defensive forecasting: the regret term scales as the norm of F (rather than its square root).

Other methods

In this subsection I will briefly list some other methods that have been used in competitive on-line regression. It appears that the benchmark classes used have always belonged to types I or III in the Kolmogorov–Tikhomirov classification. This does not mean, however, that the available prediction algorithms can be clearly divided into two groups corresponding to types I and III: quite often an ostensibly type I algorithm can be easily extended to benchmark classes that are infinite-dimensional Hilbert spaces (of type III) using the so-called "kernel trick" ([40], [35]). Sometimes the possibility of such an extension is only stated (more or less precisely) without the actual extension being carried out. In this subsection I will also discuss results of this type (which might involve some conditions of regularity that have not been stated explicitly).

Perhaps the most popular method for type III benchmark classes is Gradient Descent, together with its version, Exponentiated Gradient (the pioneering paper is [11]; see also [24] and [6]). It is very efficient computationally and often gives right orders of magnitude for the regret term. As an example, Auer *et al.* ([6], Theorem 3.1) obtain, for their prediction algorithm using Gradient

Descent, the performance guarantee

$$\sum_{n=1}^{N}(y_n-\mu_n)^2 \le \sum_{n=1}^{N}(y_n-F(x_n))^2 + 8\mathbf{c}_{\mathcal{F}}^2c^2 + 8\mathbf{c}_{\mathcal{F}}c\sqrt{\frac{1}{2}\sum_{n=1}^{N}(y_n-F(x_n))^2+\mathbf{c}_{\mathcal{F}}^2c^2} \quad (65)$$

for all N and all $F \in cU_{\mathcal{F}}$, where $\mathcal{F}$ is a Hilbert space continuously embedded in $C(\mathbf{X})$ and c is a known upper bound on $\|F\|_{\mathcal{F}}$. The regret term is bounded above by

$$8\mathbf{c}_{\mathcal{F}}^2c^2 + 8\mathbf{c}_{\mathcal{F}}c\sqrt{2N+\mathbf{c}_{\mathcal{F}}^2c^2},$$

and so its growth rate is $O(N^{1/2})$; this is typical for all popular methods for type III benchmark classes. For comparison, (57) holds with 40 replaced by 2 when $p=2$ ([44], Theorem 1).

Bounds involving the loss of the competitor in place of N, such as (65), have a clear advantage in situations where some competitors perform very well. Such bounds can also be obtained using defensive forecasting (see [44], Theorem 2).

AAR can also be carried over to Hilbert spaces (with the crucial step made in [21]). It gives a performance bound similar to (57) with $p=2$, but $40\sqrt{\mathbf{c}_{\mathcal{F}}^2+1}$ replaced by $2\mathbf{c}_{\mathcal{F}}$ ([44], Theorem 3). A simple example adapted from [11] shows that the leading constant $2\mathbf{c}_{\mathcal{F}}$ cannot be decreased further ([44], Theorem 4). (In general, attention to the constants is a tradition in learning theory that distinguishes it from some parts of the theory of function spaces; probably the impetus is coming from experimental machine learning with its common struggle for small improvements in the performance of prediction algorithms.)

7 Very big classes

In this short section we will see an example of a very fast growth rate of the regret term, barely below the useless rate of N. This slow rate is achieved not because of the richness of the function class $\mathcal{F}$ (as in §6 as compared to §5) but because of the richness of the signal space $\mathbf{X}$ itself.

The corollary of this section is rather specialized.

Corollary 9 *Suppose* $\mathbf{X}$ *is a totally bounded metric space and* γ *is a positive number that satisfy*

$$\mathcal{H}_\epsilon(\mathbf{X}) \asymp (1/\epsilon)^\gamma, \quad \epsilon \to 0. \quad (66)$$

Let $\mathcal{F} \subseteq C(\mathbf{X})$ *consist of the Hölder continuous functions of order* $\beta \in (0,1]$ *with coefficient* $c>0$ *that are bounded in absolute value by a given constant. There exists a strategy for Predictor that guarantees, for all* $F \in \mathcal{F}$*,*

$$\sum_{n=1}^{N}(y_n-\mu_n)^2 \le \sum_{n=1}^{N}(y_n-F(x_n))^2 + C_{\beta,\gamma}cN/\log^{\beta/\gamma}N \quad (67)$$

from some N on, where $C_{\beta,\gamma}$ is a constant depending only on β and γ.

Proof The modulus of continuity of F is $\omega(\epsilon) \le c\epsilon^\beta$, and so $\omega^{-1}(\epsilon) \ge (\epsilon/c)^{1/\beta}$. Substituting this in Theorem XXV (more precisely, (233)) of [27], we have

$$\log \mathcal{H}_\epsilon(\mathcal{F}) = O\left(\mathcal{H}_{(\epsilon/2c)^{1/\beta}/2}(\mathbf{X})\right),$$

which in combination with (66) gives, for small enough $\epsilon > 0$,

$$\mathcal{H}_\epsilon(\mathcal{F}) \le 2^{C_{\beta,\gamma}(c/\epsilon)^{\gamma/\beta}}.$$

Let $f(\epsilon)$ be the right-hand side of the last inequality. To estimate the infimum in (6), we find ϵ from

$$2^{C_{\beta,\gamma}(c/\epsilon)^{\gamma/\beta}} = N^{1/2}$$

(taking N instead of $N^{1/2}$ would not improve ϵ by more than a constant factor), which gives

$$\epsilon = c\left(\frac{C_{\beta,\gamma}}{\frac{1}{2}\log N}\right)^{\beta/\gamma}$$

and the upper bound

$$C'_{\beta,\gamma} cN \left(\frac{1}{\log N}\right)^{\beta/\gamma}$$

for the infimum in (6), where $C'_{\beta,\gamma}$ is another constant depending only on β and γ. ∎

In view of (38) on p. 19 we can take $\mathbf{X}$ to be the class of real-valued functions on a parallelepiped in a Euclidean space that are bounded in absolute value by a given constant and whose kth partial derivatives exist and are all Hölder continuous of order α with a given coefficient. The signal space $\mathbf{X}$ can now be interpreted as the set of images (admittedly, not very good images, without sharp boundaries between different objects).

8 The role of the norm

Our Theorems 2–4 in §3 cover all values of N, but starting from §4 we switched to stating inequalities that hold from some N on. This allowed us to simplify the statements and to tune our bounds to various parameters of the considered benchmark classes. On the negative side, however, some important information was lost: for example, the inequality (29) does not involve the norm $\|F\|_{\mathcal{F}}$ of F (whereas (49) retains the information about the norm). The reason is that asymptotically, as $N \to \infty$, the effect of $\|F\|_{\mathcal{F}}$ becomes negligible. This is only true, however, if we fix F while letting $N \to \infty$, and it can be argued that this is not the only interesting asymptotics. For example, in the experimental machine learning, N is often a constant (the size of the given data set) and it is the norm $\|F\|_{\mathcal{F}}$ of the contemplated prediction rule F that varies. Another example will

be provided by the considerations of the next section, where the norm will be chosen as a function of N. In this section we will discuss what happens if all (or all but one) values of N are taken into account. Interestingly, this will change significantly our comparative evaluation of virtues of some methods.

Finite-dimensional benchmark classes

Instead of Corollary 2 we now have:

Corollary 2* *Let $\mathcal{F}$ be a finite-dimensional Banach space embedded in $C(\mathbf{X})$ and $L \geq 1$ be a number such that*

$$\mathcal{H}_\epsilon(U_\mathcal{F}) \leq L \log \frac{1}{\epsilon}$$

for all $\epsilon \in (0, 1/2]$. There exists a strategy for Predictor that guarantees, for all $N = 2, 3, \ldots$ and all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + CL \left(\log^+ \|F\|_\mathcal{F} + \log N \right), \tag{68}$$

where C is a universal constant.

Proof Substituting $\epsilon := 1/N$ in (8), we obtain:

$$\begin{aligned}
&\sum_{n=1}^{N} (y_n - \mu_n)^2 \\
&\quad \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(L \log(\phi N) + \log\log N + \log\log \phi + 2 \right) \\
&\quad \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(2L \log \phi + 2L \log N + 2 \right),
\end{aligned}$$

which gives (68) (for a different C). ∎

Instead of Corollary 3:

Corollary 3* *Suppose $\mathbf{X}$ is a bounded set in $\mathbb{R}^m$ and $X_2 m \geq 1$. There exists a strategy for Predictor that guarantees, for all $N = 2, 3, \ldots$ and all $\theta \in \mathbb{R}^m$,*

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - \langle \theta, x_n \rangle)^2 + CX_2 m \left(\log^+ \|\theta\|_2 + \log N \right), \tag{69}$$

where C is a universal constant.

The coefficients in front of $\log N$ in the bounds (22) and (69) are not so different, m vs. $X_2 m$ (ignoring the multiplicative constants). The dependence on $\|\theta\|_2$ is, however, very different: $\|\theta\|_2^2$ vs. $X_2 m \log^+ \|\theta\|_2$, quadratic in (22) and logarithmic in (69). The explanation is that AAR uses a Gaussian prior, and so the weights assigned to remote θ decay very fast, whereas in the method of metric entropy we used slowly decaying weights. The quadratic dependence on $\|\theta\|_2$ is the price that AAR pays for computational efficiency (the former can be improved if AAR for different values of a are AA mixed, as in [44], §8, but the latter might suffer).

We can see that the relation between the AAR bound and the bound obtained using metric entropy is not as straightforward as it seemed in §4. In fact, the bounds are incomparable: among the advantages of (22) are its explicitness, a better coefficient in front of $\log N$, and the simplicity and efficiency of the underlying prediction strategy; however, the dependence of (69) on the norm of the competitor θ is better.

Benchmark classes of analytic functions

In this subsection we will be using the conventions of §5; in particular, $C(\mathbf{X})$ will be the class of continuous complex-valued functions on $\mathbf{X}$. Instead of Corollary 5 we have:

Corollary 5* *Let $\mathcal{F}$ be a Banach function space compactly embedded in $C(\mathbf{X})$ and $L, M \in [1, \infty)$ be numbers such that*

$$\mathcal{H}_\epsilon(U_\mathcal{F}) \le L \log^M \frac{1}{\epsilon} \tag{70}$$

for all $\epsilon \in (0, 1/2]$. There exists a strategy for Predictor that guarantees, for all $N = 2, 3, \dots$ and all $F \in \mathcal{F}$,

$$\sum_{n=1}^N |y_n - \mu_n|^2 \le \sum_{n=1}^N |y_n - F(x_n)|^2 + C_M L \left(\log^+ \|F\|_\mathcal{F} + \log N\right)^M, \tag{71}$$

where C_M is a constant depending only on M.

Proof Substituting $\epsilon := 1/N$ in the complex version of (8),

$$\begin{aligned}
&\sum_{n=1}^N |y_n - \mu_n|^2 \\
&\quad \le \sum_{n=1}^N |y_n - F(x_n)|^2 + C \left(L \log^M(\phi N) + \log\log N + \log\log\phi + 2\right) \\
&\quad \le \sum_{n=1}^N |y_n - F(x_n)|^2 + C' L \left(\log\phi + \log N\right)^M,
\end{aligned}$$

where C' is another universal constant. ∎

Using Corollary 5* instead of Corollary 5 gives a strategy for Predictor guaranteeing, instead of (33),

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C_{G,K} \left(\log^+ \|F\|_{A_G^K} + \log N \right)^2 \tag{72}$$

for all $F \in A_G^K$ and all $N = 2, 3, \ldots$, where $C_{G,K}$ is a constant depending on G and K only. Similarly, instead of (36) we have

$$\sum_{n=1}^{N} |y_n - \mu_n|^2 \le \sum_{n=1}^{N} |y_n - F(x_n)|^2 + C_h \left(\log^+ \|F\|_{A_h} + \log N \right)^2 \tag{73}$$

for all $F \in A_h$ and all $N = 2, 3, \ldots$, where C_h is a constant depending on h only. Notice that the asymptotic expressions (32) and (35) *per se* do not provide any information on the dependence of $C_{G,K}$ on G and K and the dependence of C_h on h.

Sobolev-type classes

Instead of Corollary 8 we now have:

Corollary 8* *Let $\mathcal{F}$ be a Banach function space compactly embedded in $C(\mathbf{X})$ and $L \ge 1$, $\gamma > 0$ be numbers satisfying*

$$\mathcal{H}_\epsilon(U_\mathcal{F}) \le L(1/\epsilon)^\gamma \tag{74}$$

for all $\epsilon \in (0, 1/2]$. There exists a strategy for Predictor that guarantees, for all $N = 1, 2, \ldots$ and all $F \in \mathcal{F}$,

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \left(L^{\frac{1}{\gamma+1}} \phi^{\frac{\gamma}{\gamma+1}} N^{\frac{\gamma}{\gamma+1}} + \log^+ \log \frac{N}{\gamma} + \log\log\phi \right), \tag{75}$$

where C is a universal constant and ϕ is defined by (9).

Proof See the proof of Corollary 8 (except that we cannot longer ignore the $\log\log$ terms in (8)). The only case that remains to be considered is where N is so small that ϵ in (41) (with L replaced by $L\phi^\gamma$) fails to belong to $(0, 1/2]$. In this case, however, the regret term of (75) exceeds $N/2$ because of the term ϵN in (8), and so we can take $C := 2$. ∎

We will refrain from stating the non-asymptotic versions of the inequalities derived in §6 for specific function classes: such versions would be awkward and would add little to our understanding of the dependence of the regret term on the competitor's norm.

9 Super-universal prediction?

In §§5–6 we dealt with universal prediction in the following, somewhat vague (as most of our informal discussion in this section), sense: for a wide (in any case, dense in $C(\mathbf{X})$) class $\mathcal{F}$ of continuous prediction rules find a prediction strategy competitive with all $F \in \mathcal{F}$. Possible dense classes $\mathcal{F}$ can be of very different size even when defined on the same domain $\mathbf{X}$. Even such meagre (barely infinite-dimensional from the point of view of metric entropy) function classes as A_G^K and A_h of §5 are dense (and so lead to a universal prediction strategy, in the sense of Theorem 1). The classes of §6 are much larger. However, we never know in advance which class $\mathcal{F}$ will work best for our data sequence $x_1, y_1, x_2, y_2, \ldots$; it would be ideal to have a prediction strategy that works well for many different $\mathcal{F}$ simultaneously. The study of existence of such "super-universal" prediction strategies is a vast understudied (and ill-defined) area, and in this section I will only make several simple and random observations.

Remark There is a cheap way of achieving "super-universality": we can AA mix prediction strategies corresponding to many different classes $\mathcal{F}$. This would, however, further impair computational efficiency and possibly lead to cumbersome performance guarantees (remember that the classes we are interested in, such as A_h, A_G^K, $B_{p,q}^s$, often depend on one or more parameters).

Let us say that a function class $\mathcal{F}_1 \subseteq C(\mathbf{X})$ "dominates" a function class $\mathcal{F}_2 \subseteq C(\mathbf{X})$ if any prediction strategy that performs not much worse than the best small-norm prediction rules in $\mathcal{F}_1$ automatically performs not much worse than the best small-norm prediction rules in $\mathcal{F}_2$. (The corresponding definition for the case where $\mathcal{F}_1$ and $\mathcal{F}_2$ are compact classes of functions is simpler: we can ignore the "small-norm" qualification.) We will not try to formalize this "definition" in this paper.

Since we do not have any lower bounds in this paper, when discussing the relation of domination we will be comparing the available performance guarantees rather than the optimal ones. Hopefully, this will be corrected in the future work.

Ideally, there would be one or very few classes $\mathcal{F}$ that would dominate numerous other natural classes. We will see in this section that less massive classes often dominate more massive ones (of course, with all the qualifications mentioned above). There is no hope for finite-dimensional classes to dominate infinite-dimensional ones, and in the three subsections of this section we will discuss the relation of domination between type II classes and between type III classes, and to what degree type II can dominate type III.

Domination between some classes of analytic functions

In this and following subsections, unlike §5, we will consider periodic period 2π real-valued functions on $\mathbb{R}$; the function space A_h is now defined as the class of all such functions that can be analytically continued to $\{z \mid |\operatorname{Im} z| \le h\}$, with

the norm defined to be the supremum norm of the analytic continuation (which is unique).

We will be interested in the quality of competition with prediction rules in A_h achieved by the prediction strategy designed for competing with prediction rules in A_H for $H > h$. But first we prove an auxiliary result.

Lemma 2 *Let $0 < h < H < \infty$ and let $F \in A_h$. For small enough $\epsilon > 0$,*

$$\log \mathcal{A}_\epsilon^{A_H}(F) \le C \frac{H}{h} \log \frac{1}{\epsilon}, \tag{76}$$

where C is a universal constant.

Proof According to Achieser's theorem ([38], 5.7.21; [1], §94) for sufficiently large J there is a trigonometric polynomial of degree J at a uniform distance from F at most

$$\frac{8c}{\pi} e^{-hJ}$$

where $c := \|F\|_{A_h}$. To make sure that this does not exceed $\epsilon > 0$ (assumed sufficiently small), it suffices to set

$$J := \left\lceil \frac{1}{h} \ln \frac{8c}{\pi\epsilon} \right\rceil. \tag{77}$$

The absolute value of the approximating trigonometric polynomial does not exceed $\|F\|_{C(\mathbb{R})} + \epsilon$ on the real line and so does not exceed

$$\left(\|F\|_{C(\mathbb{R})} + \epsilon \right) e^{JH} \tag{78}$$

in the strip $|\operatorname{Im} z| < H$ (this follows from the Phragmén–Lindelöf theorem: see [38], p. 13, footnote ****). Substituting (77) into (78), we find

$$\log \mathcal{A}_\epsilon^{A_H}(F) \le \log \left(\|F\|_{C(\mathbb{R})} + \epsilon \right) + (\log e) JH \le C \frac{H}{h} \log \frac{1}{\epsilon}$$

for small enough $\epsilon > 0$. ∎

Combining Lemma 2 with Theorem 4, we obtain the following corollary.

Corollary 10 *Let $0 < h < H < \infty$. The strategy for Predictor constructed in §5 for the benchmark class A_H guarantees*

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \le \sum_{n=1}^{N} (y_n - F(x_n))^2 + C \frac{H^2}{h^3} \log^2 N \tag{79}$$

for each $F \in A_h$ from some N on, where C is a universal constant.

Proof The regret term in (30) can be bounded above by

$$C'_M \left[L \left(\frac{H}{h} \log \frac{1}{\epsilon} \right)^M + L \log^2 \frac{1}{\epsilon} + \epsilon N \right]_{\epsilon=1/N} \\ \le C \left[\frac{H^2}{h^3} \log^2 \frac{1}{\epsilon} + \epsilon N \right]_{\epsilon=1/N} \le C' \frac{H^2}{h^3} \log^2 N.$$

In this chain, we set $M := 2$ and $L := 1/h$ (cf. (35)). ∎

The regret term in (79) is not quite as good as the regret term $\frac{C}{h} \log^2 N$ that would be obtained if we used the right value h instead of using H (cf. (36)), but the difference is not great.

From classes of type II to classes of type III

In this subsection we will see how well prediction strategies designed for type II classes can cope with type III classes. The difference between the sizes of the classes of different types is huge, and the leap might lead to losing half of the smoothness of type III classes.

Lemma 3 *Let $h > 0$ and let $F : \mathbb{R} \to \mathbb{R}$ be a non-zero periodic function with period 2π whose kth derivative ($k \in \{0, 1, \ldots\}$) exists and is Hölder continuous of order $\alpha \in (0, 1]$ with coefficient c. Set $s := k + \alpha$. For small enough $\epsilon > 0$,*

$$\log \mathcal{A}_\epsilon^{A_h}(F) \le Ch \left(\frac{12c}{\epsilon} \right)^{1/s}, \tag{80}$$

where C is a universal constant.

Proof We will emulate the proof of Lemma 2. According to Jackson's theorem ([29], Theorem 2 in §IV.3) there is a trigonometric polynomial of degree J at a uniform distance from F at most

$$\frac{12^{k+1} c (1/J)^\alpha}{J^k} = 12^{k+1} c J^{-s}.$$

This distance will not exceed $\epsilon > 0$ if we set

$$J := \left\lceil \left(\frac{12^{k+1} c}{\epsilon} \right)^{1/s} \right\rceil. \tag{81}$$

The absolute value of the approximating trigonometric polynomial does not exceed

$$\left(\|F\|_{C(\mathbb{R})} + \epsilon \right) e^{Jh} \tag{82}$$

(cf. (78)) in the strip $|\operatorname{Im} z| < h$, and so we can substitute (81) into (82) to find

$$\log \mathcal{A}_\epsilon^{A_h}(F) \le \log \left(\|F\|_{C(\mathbb{R})} + \epsilon \right) + (\log e) h \left\lceil \left(\frac{12^{k+1} c}{\epsilon} \right)^{1/s} \right\rceil,$$

which for small enough ϵ gives (80) with any $C > 12 \log e$. ∎

Remark Another way of deriving an estimate for $\mathcal{A}_\epsilon^{A_h}(F)$ for a smooth F would be to combine Kolmogorov's estimate [26] of the remainder of the Fourier series with the known results about the size of coefficients in Fourier series ([8], §I.24). This would, however, produce a weaker result.

Combining Lemma 3 with Theorem 4, we now obtain:

Corollary 11 *Let $F : \mathbb{R} \to \mathbb{R}$ be a periodic period 2π function whose kth derivative ($k \geq 0$) is Hölder continuous of order α with coefficient c. The strategy for Predictor constructed for the class A_h guarantees*

$$\sum_{n=1}^{N} (y_n - \mu_n)^2 \leq \sum_{n=1}^{N} (y_n - F(x_n))^2 + Ch^{\frac{s}{s+2}} c^{\frac{2}{s+2}} N^{\frac{2}{s+2}} \tag{83}$$

from some N on, where $s := k + \alpha$ and C is a universal constant.

Proof The proof is similar to that of Corollary 10. The regret term in (30) can be bounded above by

$$C'_M \inf_{\epsilon \in (0,1]} \left(Lh^M \left(\frac{12c}{\epsilon} \right)^{M/s} + \epsilon N \right) \tag{84}$$

(it is clear that the term $L \log^M \frac{1}{\epsilon}$ can be ignored). Using the upper bound (47) for (84), we obtain

$$2C'_M L^{\frac{s}{s+M}} h^{\frac{Ms}{s+M}} (12c)^{\frac{M}{s+M}} N^{\frac{M}{s+M}}.$$

Ignoring $12^{M/(s+M)} \in (1, 12)$ and substituting $M := 2$ and $L := 1/h$ (cf. (35)), we reduce this to (83). ∎

The growth rate $N^{2/(s+2)} = N^{1/(s/2+1)}$ of the regret term in (83) is worse than the rate $N^{1/(s+1)}$ obtained in §6 (see (50)) for a prediction strategy designed specifically for functions with Hölder continuous derivatives. We can say that one loses half of the smoothness of F when using the wrong benchmark class.

Domination between Sobolev-type classes

We first state a trivial corollary of the definition of real interpolation in terms of the K-method (in the form of the "approximation theorem" in [2], 5.31–5.32). For the definition of the K-method, see, e.g., [9], §3.1, or [2], 7.8–7.10; the notation $(X_0, X_1)_{\theta,q}$ below can be understood to be the abbreviation for $(X_0, X_1)_{\theta,q,K}$. We will be mostly interested in the case $q = \infty$.

Lemma 4 *Let (X_0, X_1) be an interpolation pair and $\theta \in (0, 1)$; set $X := (X_0, X_1)_{\theta,\infty}$. For each $F \in X$ and each $t > 0$ there exists $F_t \in X_1$ such that*

$$\begin{cases} \|F - F_t\|_{X_0} \leq 2t^\theta \|F\|_X \\ \|F_t\|_{X_1} \leq 2t^{\theta-1} \|F\|_X . \end{cases} \tag{85}$$

Proof Since the function

$$K(t,F) := \inf_{F_0 \in X_0, F_1 \in X_1 : F = F_0 + F_1} \left(\|F_0\|_{X_0} + t \|F_1\|_{X_1} \right)$$

(this is a generalization of (14)) is continuous in t ([9], Lemma 3.1.1), we have, by the definition of the K-method:

$$\begin{aligned} \|F\|_X &= \sup_{t \in (0,\infty)} t^{-\theta} K(t,F) \\ &= \sup_{t \in (0,\infty)} \inf \left\{ t^{-\theta} \|F_0\|_{X_0} + t^{1-\theta} \|F_1\|_{X_1} \;\middle|\; F = F_0 + F_1, F_0 \in X_0, F_1 \in X_1 \right\}. \end{aligned}$$

Therefore, for each $t > 0$ there is a split $F = F_0 + F_1$ such that

$$t^{-\theta} \|F_0\|_{X_0} + t^{1-\theta} \|F_1\|_{X_1} \leq 2 \|F\|_X ,$$

which is stronger than the statement of the lemma. ∎

By [9], Theorem 6.4.5(1),

$$s_0 \neq s_1 \implies \left(B^{s_0}_{p,q_0}, B^{s_1}_{p,q_1} \right)_{\theta,r} = B^{(1-\theta)s_0 + \theta s_1}_{p,r}, \tag{86}$$

and applying this to the Hölder–Zygmund spaces $\mathscr{C}^s(\mathbf{X}) := B^s_{\infty,\infty}(\mathbf{X})$ we obtain the following corollary of Lemma 4.

Corollary 12 *Let $0 < s < S < \infty$. For each $F \in \mathscr{C}^s(\mathbf{X})$ and each $\epsilon > 0$ there exists $F_\epsilon \in \mathscr{C}^S(\mathbf{X})$ such that*

$$\begin{cases} \|F - F_\epsilon\|_{C(\mathbf{X})} \leq C \epsilon^s \|F\|_{\mathscr{C}^s(\mathbf{X})} \\ \|F_\epsilon\|_{\mathscr{C}^S(\mathbf{X})} \leq 2 \epsilon^{s-S} \|F\|_{\mathscr{C}^s(\mathbf{X})} , \end{cases}$$

where C is a universal constant.

Proof Setting $\theta := s/S$, we obtain from (86):

$$\left(B^0_{\infty,1}, B^S_{\infty,\infty} \right)_{s/S,\infty} = B^s_{\infty,\infty}.$$

Remember that there is a continuous embedding $B^0_{\infty,1}(\mathbf{X}) \hookrightarrow C(\mathbf{X})$ ([18], (2.3.3/3)). It remains to set $t := \epsilon^S$ in (85). ∎

Let us apply the last corollary to the case of the performance bound (63). That bound (together with its derivation, involving the $\mathscr{C}^s(\mathbf{X})$ norm) gives the regret term of order, approximately,

$$\|F\|_{\mathscr{C}^s(\mathbf{X})} N^{1-s/m} \tag{87}$$

for the benchmark class $\mathscr{C}^s(\mathbf{X})$, $0 < s \leq m/2$, and of order

$$\|F\|_{\mathscr{C}^S(\mathbf{X})} N^{1-S/m} \tag{88}$$

for the benchmark class $\mathscr{C}^S(\mathbf{X})$, $0 < S \le m/2$. Suppose $s < S$ and let us see when a prediction strategy ensuring regret term (88) for $\mathscr{C}^S(\mathbf{X})$ automatically ensures regret term (87) for $\mathscr{C}^s(\mathbf{X})$.

Corollary 12 guarantees that every prediction strategy ensuring regret term (88) for $F \in \mathscr{C}^S(\mathbf{X})$ ensures regret term

$$\inf_{\epsilon>0} \left(\|F_\epsilon\|_{\mathscr{C}^S(\mathbf{X})} N^{1-S/m} + \|F - F_\epsilon\|_{C(\mathbf{X})} N \right) \\ \le \inf_{\epsilon>0} \left(2 \|F\|_{\mathscr{C}^s(\mathbf{X})} N^{1-S/m} \epsilon^{s-S} + C \|F\|_{\mathscr{C}^s(\mathbf{X})} N \epsilon^s \right) \quad (89)$$

for $F \in \mathscr{C}^s(\mathbf{X})$. Using the upper bound (47) for (89), we obtain regret

$$2 \left(2 \|F\|_{\mathscr{C}^s(\mathbf{X})} N^{1-S/m} \right)^{\frac{s}{S}} \left(C \|F\|_{\mathscr{C}^s(\mathbf{X})} N \right)^{\frac{S-s}{S}},$$

which coincides, to within a constant factor, with (87).

We can see that the case $s \approx m/2$ in (63) dominates all other cases with $s \le m/2$. The bound for the case $s \approx m/2$ was derived in [44] using Hilbert-space methods (applicable when $p = 2$). The Banach-space methods developed in [45] might eventually turn out to be less important (but remember that we only considered Besov spaces $B^s_{p,q}$ with p and q set to infinity).

To extend this analysis to the case $s > m/2$, we will have to compare regret terms of order

$$\|F\|^{\frac{m}{m+s}}_{\mathscr{C}^s(\mathbf{X})} N^{\frac{m}{m+s}} \quad (90)$$

for the benchmark class $\mathscr{C}^s(\mathbf{X})$ and

$$\|F\|^{\frac{m}{m+S}}_{\mathscr{C}^S(\mathbf{X})} N^{\frac{m}{m+S}} \quad (91)$$

for $\mathscr{C}^S(\mathbf{X})$, where $0 < s < S$ (see (52) with p and q set to ∞, as in the case of (50)). Since our comparison is informal anyway, we will ignore the $\log \log$ terms in (75). Corollary 12 and the upper bound (47) imply that every prediction strategy ensuring regret term (91) for $\mathscr{C}^S(\mathbf{X})$ will also ensure regret term

$$\inf_{\epsilon>0} \left(\|F_\epsilon\|^{\frac{m}{m+S}}_{\mathscr{C}^S(\mathbf{X})} N^{\frac{m}{m+S}} + \|F - F_\epsilon\|_{C(\mathbf{X})} N \right) \\ \le \inf_{\epsilon>0} \left(2 \|F\|^{\frac{m}{m+S}}_{\mathscr{C}^s(\mathbf{X})} N^{\frac{m}{m+S}} \epsilon^{(s-S)\frac{m}{m+S}} + C \|F\|_{\mathscr{C}^s(\mathbf{X})} N \epsilon^s \right) \\ \le C' \left(\|F\|^{\frac{m}{m+S}}_{\mathscr{C}^s(\mathbf{X})} N^{\frac{m}{m+S}} \right)^{\frac{s(m+S)}{S(m+s)}} \left(\|F\|_{\mathscr{C}^s(\mathbf{X})} N \right)^{\frac{(S-s)m}{S(m+s)}} \\ = C' \left(\|F\|_{\mathscr{C}^s(\mathbf{X})} N \right)^{\frac{m}{m+s}} \quad (92)$$

for $\mathscr{C}^s(\mathbf{X})$. The regret rate obtained is as good as (90), to within a constant factor. Therefore, as far as our bounds are concerned, $\mathscr{C}^S(\mathbf{X})$ dominates $\mathscr{C}^s(\mathbf{X})$. Unfortunately, these bounds are known to be loose (see §6), at least in the case of low smoothness, and it remains to be seen whether the domination still holds for tighter bounds.

10 Conclusion

In this paper we have seen the following typical rates of growth of the regret term:

(I) for finite dimensional $\mathcal{F}$ (type I of [27], §3),

$$O(\log N);$$

(II) for classes $\mathcal{F}$ of analytic functions of m variables (type II of [27]),

$$O\left(\log^{m+1} N\right);$$

(III) for classes $\mathcal{F}$ of functions of m variables with smoothness indicator s (type III of [27]),

$$O\left(N^{\frac{m}{m+s}}\right);$$

(IV) for classes $\mathcal{F}$ of Lipschitzian functionals on classes of the previous type (such $\mathcal{F}$ are representative of type IV of [27]), a typical rate is

$$O\left(N/\log^{s/m} N\right).$$

Rates of types I and III have been known in competitive on-line prediction, whereas types II and IV appear new. For the first time we can see enough fragments to get an impression of the big picture. These are still small fragments and the picture is still vague. The method of metric entropy, despite its wide applicability, is not universal and often does not give optimal results. My goal was to convince my listeners or readers that the arising questions are interesting ones.

We have also considered, in a very tentative way, the question of how much one has to pay for using a wrong benchmark class (§9). From the available very preliminary results it appears that using meagre (albeit dense in $C(\mathbf{X})$) benchmark classes is safer than using rich classes.

These are possible directions of theoretical research:

- Find computationally efficient prediction strategies for benchmark classes such as A_G^K and A_h (type II) and Besov spaces with $m/(m+s) < 1/2$ (in the notation of (52)).
- Find uniform in N estimates of metric entropy (for applications such as those in §§8–9).
- Extend this paper's results to discontinuous prediction rules (for estimates of metric entropy in this case see, e.g., [14]).
- Perhaps most importantly, complement performance guarantees such as those in this paper with lower bounds. A lower bound corresponding to Proposition 1 with $p = 2$ is proved in [44], Theorem 4; however, the function space $\mathcal{F}$ constructed there is not compactly embedded in $C(\mathbf{X})$, and so not interesting from the point of view of metric entropy.

In experimental research, it would be interesting to find out the "empirical approachability function"

$$\mathcal{A}_\epsilon^{\mathcal{F},2}(x_1, y_1, \dots, x_N, y_N) := \inf\left\{ \|F\|_{\mathcal{F}} \;\middle|\; \frac{1}{N}\sum_{n=1}^{N}(y_n - F(x_n))^2 \le \epsilon \right\} \tag{93}$$

(cf. (12); the upper index 2 refers to using the quadratic loss function in this definition) for standard benchmark data sets

$$(x_1, y_1, \dots, x_N, y_N) := ((x_1, y_1), \dots, (x_N, y_N)) \tag{94}$$

and standard function classes $\mathcal{F}$. It is clear that (93) will be finite for all $\epsilon > 0$ if $\mathcal{F}$ is dense in $C(\mathbf{X})$ and

$$x_{n_1} = x_{n_2} \implies y_{n_1} = y_{n_2}.$$

If a prediction strategy guarantees a regret term of $f(\|F\|_{\mathcal{F}}, N)$ (we will assume that f is a continuous function in its first argument), in the sense that

$$\sum_{n=1}^{N}(y_n - \mu_n)^2 \le \sum_{n=1}^{N}(y_n - F(x_n))^2 + f(\|F\|_{\mathcal{F}}, N)$$

for all $F \in \mathcal{F}$ and all $N = 1, 2, \dots$, the loss of this prediction strategy on the data set (94) will be at most

$$\inf_{\epsilon>0}\Big(f\left(\mathcal{A}_\epsilon^{\mathcal{F},2}(x_1, y_1, \dots, x_N, y_N), N\right) + \epsilon N \Big).$$

Knowing typical empirical approachability functions (93) for various function classes might suggest function classes most promising for various practical problems.

A natural next step would be to compare different benchmark classes on real-world data sets. This is a task for experimental machine learning; what learning theory can do is to study the relation of domination between various *a priori* plausible benchmark classes: e.g., some of them may turn out to be useless or nearly useless on purely theoretical grounds.

Acknowledgments

This paper was written to support my talk at the workshop "Metric entropy and applications in analysis, learning theory and probability" (Edinburgh, Scotland, September 2006). I am grateful to its organizers, Thomas Kühn, Fernando Cobos and W. D. Evans, for inviting me. This version of the paper is preliminary and is likely to be revised as a result of discussions at the workshop.

Nicolò Cesa-Bianchi, Gábor Lugosi, Steven Smale and Alex Smola supplied the principal components of this paper with their incisive questions and comments. Ilia Nouretdinov's help was invaluable. This work was partially supported by MRC (grant S505/65).

References

[1] Naum I. Achieser. *Theory of Approximation*. Ungar, New York, 1956.

[2] Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics*. Academic Press, Amsterdam, second edition, 2003.

[3] Lars V. Ahlfors. *Complex Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, third edition, 1979.

[4] Nachman Aronszajn. La théorie générale des noyaux reproduisants et ses applications, première partie. *Proceedings of the Cambridge Philosophical Society*, 39:133–153 (additional note: p. 205), 1943. The second part of this paper is [5].

[5] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[6] Peter Auer, Nicolò Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64:48–75, 2002.

[7] V. Bargmann. On a Hilbert space of analytic functions and an associated integral transform, part 1. *Communications on Pure and Applied Mathematics*, 14:187–214, 1961.

[8] Nina K. Bary. *A Treatise on Trigonometric Series*. Macmillan, New York, 1964. In two volumes. Ralph P. Boas, Jr., is very critical of the English translation in his review in *Mathematical Reviews*. Russian edition: Bari, Nina K. Trigonometricheskie ryady. Fizmatlit, Moscow, 1961.

[9] Jöran Bergh and Jörgen Löfström. *Interpolation Spaces: An Introduction*, volume 223 of *Die Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 1976.

[10] Bernd Carl and Irmtraud Stephani. *Entropy, Compactness and the Approximation of Operators*, volume 98 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, England, 1990.

[11] Nicolò Cesa-Bianchi, Philip M. Long, and Manfred K. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

[12] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, England, 2006.

[13] James A. Clarkson. Uniformly convex spaces. *Transactions of the American Mathematical Society*, 40:396–414, 1936.

[14] G. F. Clements. Entropies of sets of functions of bounded variation. *Canadian Journal of Mathematics*, 15:422–432, 1963.

[15] Fernando Cobos and David E. Edmunds. Clarkson's inequalities, Besov spaces and Triebel–Sobolev spaces. *Zeitschrift für Analysis und ihre Anwendungen*, 7:229–232, 1988.

[16] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin (New Series) of the American Mathematical Society*, 39:1–49, 2002.

[17] Richard M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, England, revised edition, 2002.

[18] David E. Edmunds and Hans Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, England, 1996.

[19] Ryszard Engelking. *General Topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann, Berlin, second edition, 1989.

[20] Dean P. Foster. Prediction in the worst case. *Annals of Statistics*, 19:1084–1090, 1991.

[21] Alex Gammerman, Yuri Kalnishkan, and Vladimir Vovk. On-line prediction with kernels and the Complexity Approximation Principle. In Max Chickering and Joseph Halpern, editors, *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence*, pages 170–176, Arlington, VA, 2004. AUAI Press.

[22] Olof Hanner. On the uniform convexity of L^p and l^p. *Arkiv för Matematik*, 3:239–244, 1956.

[23] Yuri Kalnishkan and Michael V. Vyugin. The Weak Aggregating Algorithm and weak mixability. In Peter Auer and Ron Meir, editors, *Proceedings of the Eighteenth Annual Conference on Learning Theory*, volume 3559 of *Lecture Notes in Computer Science*, pages 188–203, Berlin, 2005. Springer.

[24] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated Gradient versus Gradient Descent for linear predictors. *Information and Computation*, 132:1–63, 1997.

[25] Jyrki Kivinen and Manfred K. Warmuth. Averaging expert predictions. In Paul Fischer and Hans U. Simon, editors, *Proceedings of the Fourth European Conference on Computational Learning Theory*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 153–167, Berlin, 1999. Springer.

[26] Andrei N. Kolmogorov. Zur Grössenordnung des Restgliedes Fourierschen Reihen differenzierbarer Functionen. *Annals of Mathematics*, 36:521–526, 1935.

[27] Andrei N. Kolmogorov and Vladimir M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces (in Russian). *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[28] Joram Lindenstrauss and Lior Tzafriri. *Classical Banach Spaces II: Function Spaces*, volume 97 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Berlin, 1979.

[29] Isidor P. Natanson. *Constructive Function Theory*, volume 1: Uniform Approximation. Ungar, New York, 1964.

[30] Vern I. Paulsen. An introduction to the theory of reproducing kernel Hilbert spaces. Course notes, available from the author's web page (accessed in August 2006), February 2006.

[31] Lev S. Pontryagin and Lev G. Shnirel'man. Sur une propriété métrique de la dimension. *Annals of Mathematics (New Series)*, 33:156–162, 1932.

[32] Walter Rudin. *Real and Complex Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, third edition, 1987.

[33] Saburou Saitoh. *Integral Transforms, Reproducing Kernels and their Applications*, volume 369 of *Pitman Research Notes in Mathematics*. Longman, Harlow, England, 1997.

[34] Craig Saunders, Mark O. Stitson, Jason Weston, Leon Bottou, Bernhard Schölkopf, and Alexander J. Smola. Support vector machine reference manual. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, 1998.

[35] Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[36] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[37] Ingo Steinwart, Don Hush, and Clint Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. Technical Report LA-UR 04-8274, Los Alamos National Laboratory, 2004.

[38] Aleksandr F. Timan. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, Oxford, 1963.

[39] Hans Triebel. *Theory of Function Spaces II*, volume 84 of *Monographs in Mathematics*. Birkhäuser, Basel, 1992.

[40] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[41] Anatoly G. Vitushkin. *Otsenka slozhnosti zadachi tabulirovaniya*. Fizmatlit, Moscow, 1959. English translation: *Theory of the Transmission and Processing of Information*, Pergamon Press, Oxford, 1961.

[42] Vladimir Vovk. Competitive on-line statistics. *International Statistical Review*, 69:213–248, 2001.

[43] Vladimir Vovk. Non-asymptotic calibration and resolution. Technical Report `arXiv:cs.LG/0506004` (version 3), `arXiv.org` e-Print archive, August 2005.

[44] Vladimir Vovk. On-line regression competitive with reproducing kernel Hilbert spaces. Technical Report `arXiv:cs.LG/0511058` (version 2), `arXiv.org` e-Print archive, January 2006.

[45] Vladimir Vovk. Competing with wild prediction rules. Technical Report `arXiv:cs.LG/0512059` (version 2), `arXiv.org` e-Print archive, January 2006.

[46] Vladimir Vovk. Predictions as statements and decisions. Technical Report `arXiv:cs.LG/0606093`, `arXiv.org` e-Print archive, June 2006.

[47] Vladimir Vovk. Competing with stationary prediction strategies. Technical Report `arXiv:cs.LG/0607067`, `arXiv.org` e-Print archive, July 2006.

[48] Vladimir Vovk. Competing with Markov prediction strategies. Technical Report `arXiv:cs.LG/0607136`, `arXiv.org` e-Print archive, July 2006.

[49] Vladimir Vovk. Leading strategies in competitive on-line prediction. Technical Report `arXiv:cs.LG/0607134`, `arXiv.org` e-Print archive, July 2006.

[50] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.

[51] Harold Widom. Rational approximation and n-dimensional diameter. *Journal of Approximation Theory*, 5:343–361, 1972.